Curso Cálculo Numérico MS 211

Prof^o: Eduardo Abreu - IMECC - DMA

UNICAMP/IMECC - 1^o Semestre 2015
http://www.ime.unicamp.br/∼ms211-cursao/

- **O curso Cálculo Numério**
- **Início Previsto pelo Calendário DAC/UNICAMP**

– Ementa, Critérios de avaliação e Referências bibliográficas

– Linguagens de programação

- O curso (Ementa, Critérios de avaliação e Referências bibliográficas)

- **Ementa** (Tópicos)

1) **Impacto da Computação em Precisão Finita**: Aritmética de Ponto Flutuante e Erros em Operações Numéricas. (Perda de dígitos significativos e Condicionamento de algoritmos). Teorema de Taylor.

2) **Zeros reais de Funções Reais** (equações não-lineares - escalar): Método da bissecção. Método de Newton. Método da Secante (um método do tipo "quase-Newton").

3) **Resolução de Sistemas Lineares** - Métodos **diretos**: Eliminação de Gauss e Decomposição LU. Métodos **iterativos**: Gauss-Jacobi e Gauss-Seidel.

4) **Resolução de Sistemas não Lineares**: Método de Newton para sistemas.

5) **Resolução numérica de equações diferenciais ordinárias**. Problemas de **valor inicial**: método de Euler, métodos de série de Taylor e de Runge-Kutta. Equações de ordem superior. Problemas de **valor de contorno**: método das diferenças finitas.

6) **Aproximação**: Ajuste de curvas via método dos quadrados mínimos (quadrados mínimos lineares e linearizáveis).

7) **Aproximação**: Interpolação polinomial. Formas de Lagrange e de Newton. Erro de interpolação. Interpolação linear por partes (Spline linear).

8) **Integração Numérica**: Fórmulas de Newton-Cotes e Quadratura Gaussiana.

Critérios de avaliação

No semestre, serão aplicadas **DUAS** provas:

- A primeira prova **P1** versará sobre os tópicos de 1 a 4 da ementa.

- Enquanto que a segunda prova **P2** abrangerá **principalmente** os tópicos de 5 a 8.

Outra avaliação **MT** (projetos, listas de exercícios, testes em sala de aula, etc.), **a critério do professor** também será utilizada para compor a nota do aluno.

http://www.ime.unicamp.br/∼ms211-cursao/

- Cálculo da média final: M = (P1 + P2 + MT)/3,

  P1 e P2 são as notas das provas 1 e 2, respectivamente, e MT é a nota do projeto.

  Se M >= 7.0 e o aluno tiver pelo menos 75% de presença, então o aluno está aprovado e dispensado do exame, sendo sua média final MF = M.

  Agora, se M < 7.0 e se a frequência nas aulas for superior a 75%, então o aluno deverá, obrigatoriamente, fazer o exame. Neste caso, sua média final será calculada como MF = (M + E)/2, onde E é a nota do exame.

  Se MF < 5.0 o aluno estará reprovado na disciplina de Cálculo Numérico.

Na realização das provas:

- **Datas das provas e do exame** ver
  http://www.ime.unicamp.br/∼ms211-cursao/

1.) É obrigatória a apresentação da identidade estudantil.
2.) O aluno deve trazer sua calculadora, não poderá usar calculadoras de aparelhos celulares.
3.) O aluno que faltar a uma das provas tem um prazo de 15 dias, a partir da data da prova, para entregar ao professor responsável da turma os documentos que justifiquem esta falta, de acordo com o artigo 72 do Regimento Geral da Graduação (UNICAMP).

- **Monitoria de Cálculo Numérico** (Horários)
- Existem monitores PAD e PED, em horários diversos
- Consultar http://www.ime.unicamp.br/∼ms211-cursao/

Referências bibliográficas

- Márcia A. Gomes Ruggiero e Vera Lúcia da Rocha Lopes, Cálculo Numérico, Pearson Education do Brasil, São Paulo, segunda edição, 2000.

- Maria Cristina Cunha, Métodos Numéricos para as Engenharias e Ciências Aplicadas, Editora da Unicamp, Campinas, segunda edição, 2003.

- Análise Numérica, R. L. Burden e J. D. Faires. Editora Pioneira, 2003.

- Numerical Computing with MatLab, Cleve B. Moler, Editora SIAM, 2004. (Capítulos 1 a 7).

Referências bibliográficas (consulta adicional)

- Arieh Iserles, A first course in the numerical analysis of differential equations. U.K.: Cambridge University Press, 2009.

- K. Atkinson, Theoretical numerical analysis: a functional analysis framework, 3rd ed, 2010.

- E. Hairer, S.P. Norsett, G. Wanner., Solving ordinary differential equations I: nonstiff problems I., ed. 2009.

- E. Hairer, S.P. Norsett, G. Wanner., Solving ordinary differential equations I: stiff differential - algebraic problems II, ed. 2010.

Referências bibliográficas (consulta adicional)

- David Kincaid e Ward Cheney, Numerical Analysis, Brooks-Cole, 1991.

- J. D., Lambert, Numerical methods for ordinary differential systems: the initial value problem, U.K.: John Wiley, 1991.

- G. Hammerlin e K.-H. Hoffmann, Numerical mathematics (translated by Larry Schumaker) Springer, 1991; New York, N.Y.: Série (**Undergraduate** texts in mathematics)

- J. W. Demmel, Applied numerical linear algebra, SIAM - Society for Industrial and Applied Mathematics, 1997.

Referências bibliográficas (consulta adicional)

- Numerical methods for special functions, Amparo Gil, Javier Segura, Nico M. Temme. Philadelphia, PA: SIAM, 2007.

- John H. Mathews e Kurtis D. Fink, Numerical Methods Using MATLAB, Pearson Prentice Hall, quarta edição, 2007.

- D. Hanselman e B. Littlefield, MATLAB 6 - Curso completo, Pearson Education do Brasil, São Paulo, segunda edição, 2003.

IMPORTANTE:

- A biblioteca do IMECC tem **TODOS** os livros indicados.
- Esses livros abrangem a ementa do curso de Cálculo Numérico[1], em diferentes níveis de profundidade e nos dois **relevantes** aspectos:
  - **teórico** (fundamentação matemática) e
  - **prático** (aplicações computacionais)

[1] O material presente neste arquivo foi elaborado com base na bibliografia anteriomente indicada para o curso de Cálculo Numérico MS211. Críticas, Sugestões ou comentários são bem-vindos, *eabreu@ime.unicamp.br*

- Linguagens de programação

A que for mais confortável para vocês

Por exemplo:

- **Interpretadas** (alto nível): Matlab, Maple, Mathematica, etc...
- **Compiladas** (baixo nível): C, C++, FORTRAN, PASCAL, etc...

- Lembrandro...

  Informações sobre o curso MS211:

  – Ementa,
  – Critérios de avaliação e
  – Referências bibliográficas,

  estão disponíveis em

  http://www.ime.unicamp.br/∼ms211-cursao/

Tópico 01

Aritmética de Ponto Flutuante e Erros em Operações Numéricas. (Perda de dígitos significativos e Condicionamento de algoritmos). Teorema de Taylor.

- **Impacto da Computação em Precisão Finita:**

– Perda de Dígitos Significativos
- Tipos de erros e incertezas
- Sistema de Numeração Utilizado pelo Computador
- Representação de números no sistema $F(\beta, t, m, M)$
- Operações aritméticas em ponto flutuante

– Condicionamento de algoritmos - efeitos numéricos
- Cancelamento
- Propagação do erro
- Instabilidade numérica
- Mal condicionamento

– Exemplos de “desastres numéricos” (ou falha humana ?)

- **Tipos de erros e incertezas**

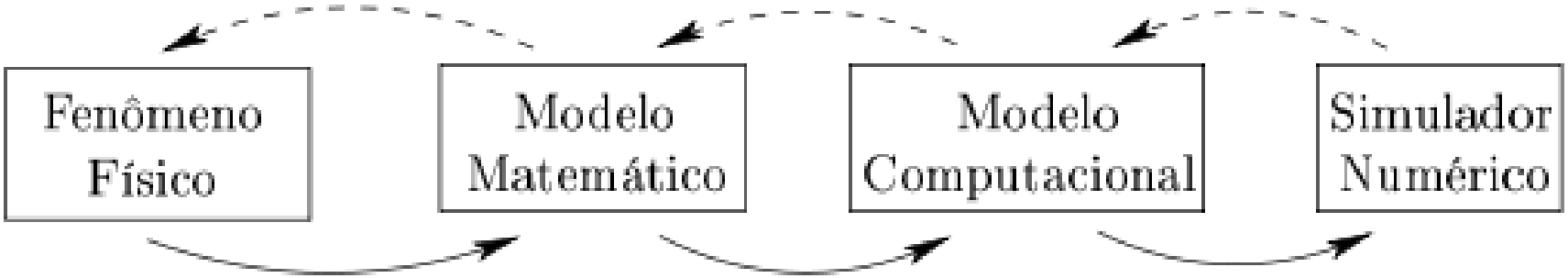


FIG 1. TIPOS DE ERROS E INCERTEZAS

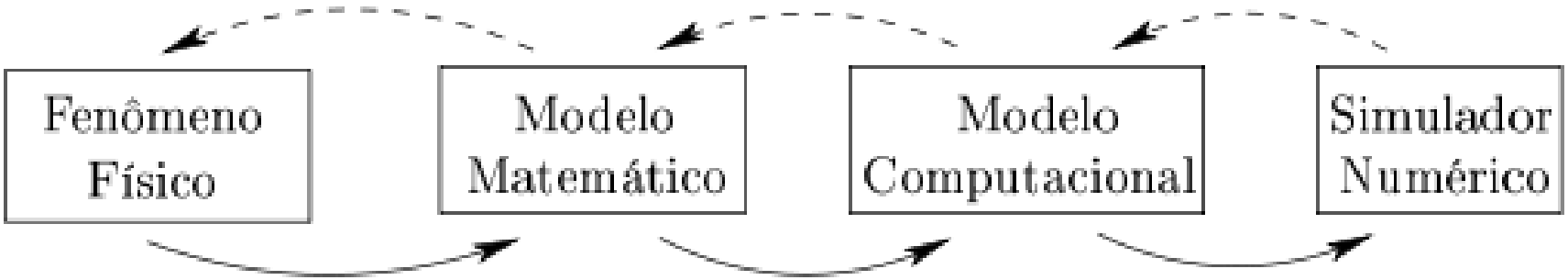


FIG 1. TIPOS DE ERROS E INCERTEZAS

- **Erro (incerteza) nos dados de entrada/medidos**

→ $574,39 \pm 0,28$ mm

→ $143,57 \pm 0,71$ ml

→ $95,27 \pm 0,46$ Kg

→ $127,00 \pm 2,54$ Volts

→ Censo populacional (altura, peso, etc...)

→ Economia de um país

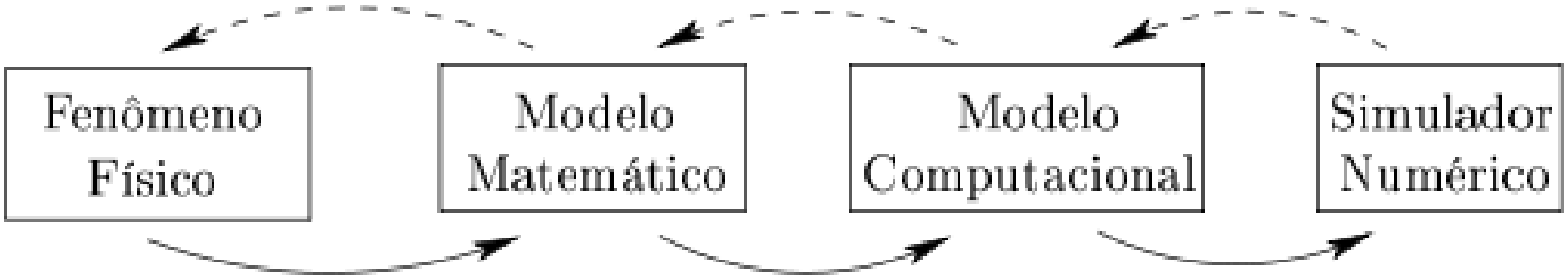


FIG 1. TIPOS DE ERROS E INCERTEZAS

- **Erro na simplificação do modelo matemático**
  EXEMPLO: Modelagem de Sistemas Complexos
  Processos na vida real são inerentemente multi-física e multi-escala (no tempo e espaço)

→ Dinâmica de fluidos
→ Astrofísica
→ Processos físico-químicos
→ Sistemas biológicos (e.g., humano)

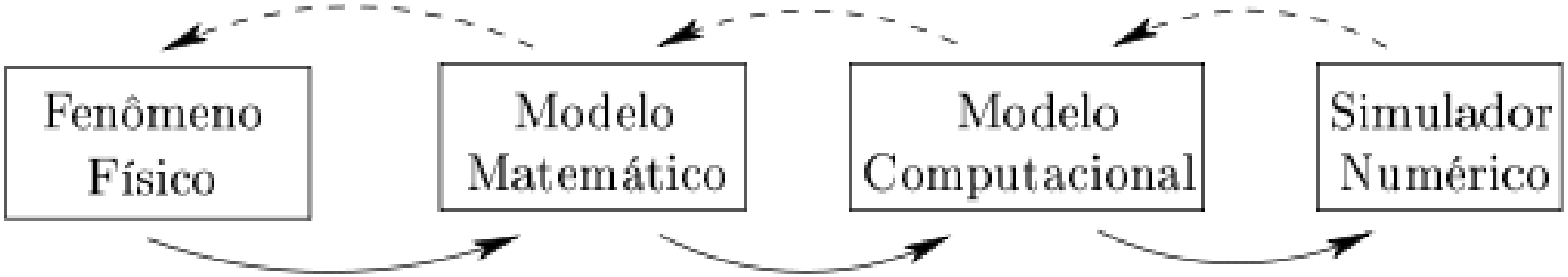


FIG 1. TIPOS DE ERROS E INCERTEZAS

- **Erro de arredondamento**
- **Erro de truncamento**

Neste curso estamos interessados nos dois últimos (**perda de dígitos significativos** no contexto de arimética de ponto flutuante), i.e., entender sua fonte, propagação, magnitude, e a taxa de crescimento e a quantificação desses erros.

- EXEMPLO (Erro de arredondamento)

- Vamos resolver, **de duas formas**, para as variáveis x e y, o seguinte sistema linear 2×2 de equações:

$$0.1036x + 0.2122y = 0.7381$$
$$0.2081x + 0.4247y = 0.9327$$

(**1**) Vamos considerar somente **três** dígitos significativos de precisão nos cálculos.

- **Lembrete**: **Dígitos significativos** são dígitos que iniciam com o dígito **não nulo** mais à esquerda e terminam com o dígito **mais correto** à direita

EXEMPLOS

$0.0035 \rightarrow 0.35 \times 10^{-2}$ tem dois dígitos significativos

$0.03017 \rightarrow 0.3017 \times 10^{-1}$ tem quatro dígitos significativos

$0.33011 \rightarrow 0.33011 \times 10^{0}$ tem cinco dígitos significativos

$0.0001469 \rightarrow 0.1469 \times 10^{-3}$ tem quatro dígitos significativos

- EXEMPLO (Erro de arredondamento)

  Procedimento: Vamos arredondar todos os números no problema **original** para **três** dígitos significativos e, a cada etapa, arredondar todos os cálculos mantendo somente **três** dígitos significativos.

$$\begin{aligned} \mathbf{L_1} &: 0.1036x + 0.2122y = 0.7381 \\ \mathbf{L_2} &: 0.2081x + 0.4247y = 0.9327 \end{aligned}$$

Fazendo o arredondamento $\cdots$

$$\begin{aligned} 0.104x + 0.212y &= 0.738 \quad L_1 \leftarrow L_1 \\ 0.208x + 0.425y &= 0.933 \quad L_2 \leftarrow L_2 \end{aligned}$$

- EXEMPLO (Erro de arredondamento)

  Procedimento: Vamos arredondar todos os números no problema **original** para **três** dígitos significativos e, a cada etapa, arredondar todos os cálculos mantendo somente **três** dígitos significativos.

$$\begin{aligned} 0.104x + 0.212y &= 0.738 \quad L_1 \leftarrow L_1 \\ 0.208x + 0.425y &= 0.933 \quad L_2 \leftarrow L_2 - \alpha L_1 \end{aligned}$$

- O multiplicador $\alpha = \dfrac{0.208}{0.104} \approx 2.00$,

$$\begin{aligned} 0.104x + 0.212y &= \quad 0.738 \\ 0.001y &= -0.547 \end{aligned}$$

- EXEMPLO (Erro de arredondamento)

  Procedimento: Vamos arredondar todos os números no problema **original** para **três** dígitos significativos e, a cada etapa, arredondar todos os cálculos mantendo somente **três** dígitos significativos.

$$\begin{aligned} 0.104x + 0.212y &= 0.738 \quad L_1 \leftarrow L_1 \\ 0.208x + 0.425y &= 0.933 \quad L_2 \leftarrow L_2 - \alpha L_1 \end{aligned}$$

- O multiplicador $\alpha = \dfrac{0.208}{0.104} \approx 2.00$,

$$\begin{aligned} 0.104x + 0.212y &= 0.738 \\ 0.001y &= -0.547 \end{aligned}$$

- Solução: $y = -\dfrac{0.547}{0.001} \approx -547$ e $x \approx 0.111 \times 10^4$.

- EXEMPLO (Erro de arredondamento)

- (2) Vamos agora repetir os mesmos cálculos com **quatro** dígitos significativos para o sistema original:

$$0.1036x + 0.2122y = 0.7381$$
$$0.2081x + 0.4247y = 0.9327$$

- EXEMPLO (Erro de arredondamento)

$$\begin{aligned} 0.1036x + 0.2122y &= 0.7381 \quad L_1 \leftarrow L_1 \\ 0.2081x + 0.4247y &= 0.9327 \quad L_2 \leftarrow L_2 - \alpha L_1 \end{aligned}$$

- O multiplicador $\alpha = \dfrac{0.2081}{0.1036} \approx 2.009$,

- EXEMPLO (Erro de arredondamento)

$$\begin{aligned} 0.1036x + 0.2122y &= 0.7381 \quad L_1 \leftarrow L_1 \\ 0.2081x + 0.4247y &= 0.9327 \quad L_2 \leftarrow L_2 - \alpha L_1 \end{aligned}$$

- O multiplicador $\alpha = \dfrac{0.2081}{0.1036} \approx 2.009$,

$$\begin{aligned} 0.1036x + 0.2122y &= 0.7381 \\ -0.0016y &= -0.5503 \end{aligned}$$

- EXEMPLO (Erro de arredondamento)

$$\begin{aligned} 0.1036x + 0.2122y &= 0.7381 \quad L_1 \leftarrow L_1 \\ 0.2081x + 0.4247y &= 0.9327 \quad L_2 \leftarrow L_2 - \alpha L_1 \end{aligned}$$

- O multiplicador $\alpha = \dfrac{0.2081}{0.1036} \approx 2.009$,

$$\begin{aligned} 0.1036x + 0.2122y &= 0.7381 \\ -0.0016y &= -0.5503 \end{aligned}$$

- Solução: $y = \dfrac{-0.5503}{-0.0016} \approx 342.9$ e $x \approx 695.2$.

  A resposta mudou de $y = -547$ para $y = 343.9$!!!

- A “pertubação” de UM dígito significativo fez MUITA diferença !!

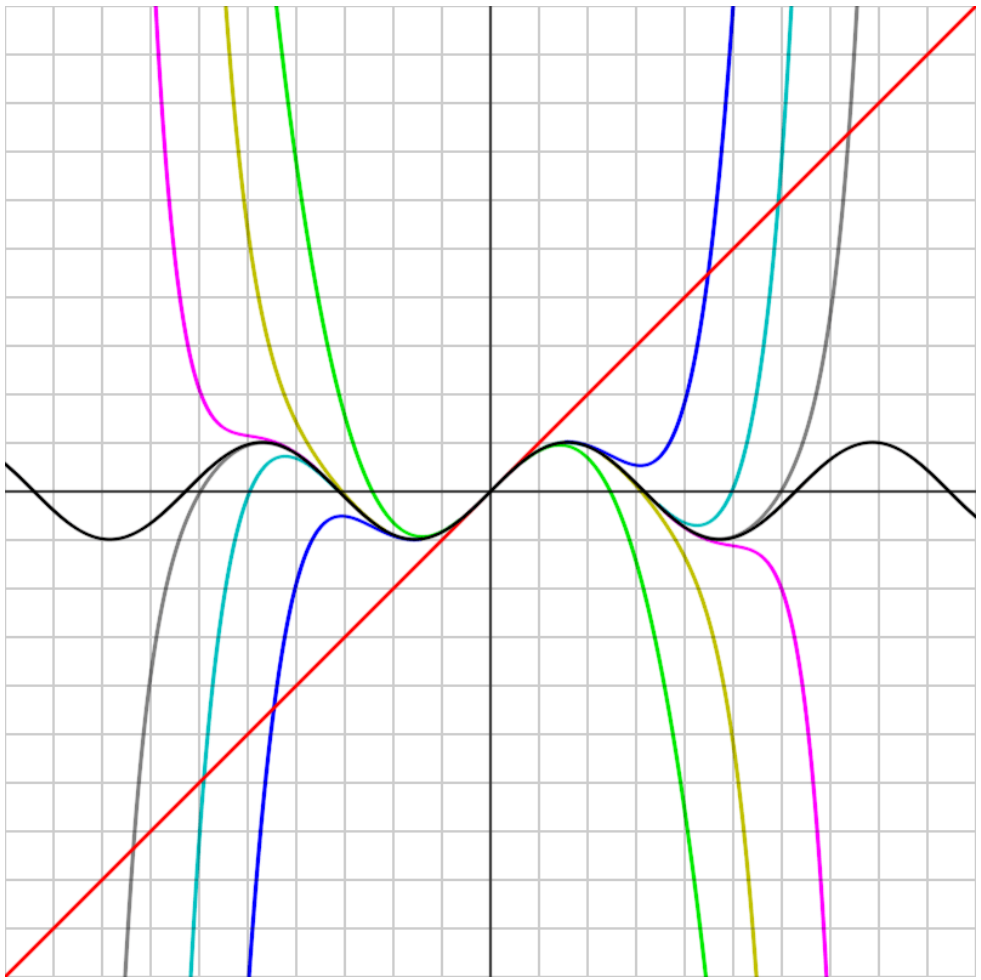
- Exemplos (Erro de truncamento)

- Cálculo de $e^{0.1}$, $sen(0.1)$ e $cos(0.1)$ usando série de Taylor:

- $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$
- $sen\, x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots$
- $cos\, x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots$

Com a ferramenta matemática “série de Taylor”, podemos escrever, por exemplo, funções **trigonométricas**, **exponenciais**, **logarítmicas** em POLINÔMIOS

Função seno de x e aproximações de série de Taylor com polinômios de grau 1, 3, 5, 7, 9, 11 e 13.



- EXEMPLOS (Erro de truncamento)

- Para o cálculo **efetivo** (na prática) de $e^{0.1}$, $sen(0.1)$ e $cos(0.1)$ precisamos *truncar* a série (uso de um número finito de termos).

- Pela **fórmula do erro** $R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-c)^{n+1}$ temos que os respectivos erros $0 \leq x \leq 0.1$ $(c = 0)$ serão inferiores a (com os dois primeiros termos):

- Erros $\frac{|x|^2}{2!}$, $\frac{|x|^5}{5!}$ e $\frac{|x|^4}{4!}$.

- Sob as hipóteses do teorema de Taylor, temos uma **fórmula explícita** para o erro cometido na aproximação !!!

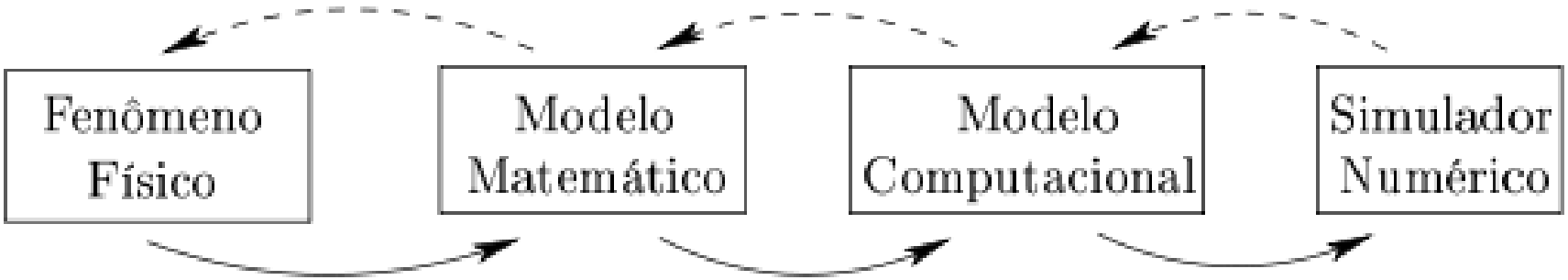


FIG 1. TIPOS DE ERROS E INCERTEZAS

- **Representação de números reais com um número fínito de dígitos significativos**

Neste curso também vamos levar em conta esse tipo de erro.

- Erro de Representação I

→ Depende da capacidade de representação numérica da máquina disponível (e do “tempo” disponível para efetuar os cálculos)

$\pi = 3.14159$

$\pi = 3.1415926535897$

$\pi = 3.14159265358979323846264338327950288 41 \cdots$

$\frac{4}{3} = 1.33333333 \cdots$

$\sqrt{3} = 1.732050807568877293527446341505872366 9 \cdots$

- Erro de Representação **II**

→ Um número pode ter uma representação finita (e precisa) em uma base e **não finita** em outra base (**isto independe da máquina utilizada !!!**)

$$(3.8)_{10} = (11.1100110\overline{01100})_2$$

$$(0.1)_{10} = (0.0001100\overline{11001})_2$$

- **Lembrete**: Regra de conversão de base

  EXEMPLO $(3.75)_{10} = (11.11)_2$

- **Lembrete**: Regra de conversão de base

  EXEMPLO $(3.75)_{10} = (11.11)_2$

- **Calcule**:

  $(1101)_2 = (\ ?\ )_{10}$
  $(0.110)_2 = (\ ?\ )_{10}$

- Para manter uma resposta precisa deve-se realizar as contas com máxima precisão ao longo dos cálculos intermediários e somente realizar algum procedimento de arredondamento no final

- É importante usar técnicas de aproximação que nos permitam **quantificar o erro** que cometemos nos cálculos (**isso é a regra em qualquer teoria de aproximação !**)

- Com o objetivo de medir os erros de **truncamento** e **arredondamento** nos cálculos numéricos, vamos discutir os **erros**:

  Absoluto, relativo e percentual.

- **Erros Absoluto, relativo e percentual**

- Suponha que x e x^* são dois números, sendo um deles uma aproximação do outro.

- O erro de x^* como uma aproximação de x é $x - x^*$.

- O **erro absoluto** de x^* como uma aproximação de x é definido por:

$$EA_x \equiv |x - x^*|$$

- Suponha que x e x^* são dois números, sendo um deles uma aproximação do outro

- O erro de x^* como uma aproximação de x é $x - x^*$
- O **erro absoluto** de x^* como uma aproximação de x é definido por:

$$EA_x \equiv |x - x^*|$$

EXEMPO: Considere os números 30.1358 e 1.1358

- Suponha que x e x^* são dois números, sendo um deles uma aproximação do outro

- O erro de x^* como uma aproximação de x é $x - x^*$
- O **erro absoluto** de x^* como uma aproximação de x é definido por:

$$EA_x \equiv |x - x^*|$$

EXEMPO: Considere os números 30.1358 e 1.1358

Sejam 30.0000 e 1, respectivamente, aproximações de 30.1358 e 1.1358

O erro absoluto de 30.0000, como uma aproximação de 30.1358, é:

$$|30.1358 - 30.0000| = 0.1358$$

O erro absoluto de 30.0000, como uma aproximação de 30.1358, é:

$$|30.1358 - 30.0000| = 0.1358$$

Note que o erro absoluto de 1, como uma aproximação de 1.1358, é:

$$|1.1358 - 1| = 0.1358$$

O erro absoluto de 30.0000, como uma aproximação de 30.1358, é:

$$|30.1358 - 30.0000| = 0.1358$$

Note que o erro absoluto de 1, como uma aproximação de 1.1358, é:

$$|1.1358 - 1| = 0.1358$$

Claramente observamos que os **erros absolutos** são os mesmos.

O erro absoluto de 30.0000, como uma aproximação de 30.1358, é:

$$|30.1358 - 30.0000| = 0.1358$$

Note que o erro absoluto de 1, como uma aproximação de 1.1358, é:

$$|1.1358 - 1| = 0.1358$$

Claramente observamos que os **erros absolutos** são os mesmos.

Pergunta: Mas, com base nos **erros absolutos**, podemos dizer que 30.0000 e 1 representam aproximações com a mesma precisão ?

Para responder a pergunta, precisamos comparar a ordem de grandeza (o peso) dos **erros absolutos** com os respectivos valores exatos (sempre quando disponíveis).

Para responder a pergunta, precisamos comparar a ordem de grandeza (o peso) dos **erros absolutos** com os respectivos valores exatos.

Vamos comparar por meio do **erro relativo**.

Para responder a pergunta, precisamos comparar a ordem de grandeza (o peso) dos **erros absolutos** com os respectivos valores exatos.

Vamos comparar por meio do **erro relativo**.

- O **erro relativo** de x^* como uma aproximação de x é defido por:

$$ER_x \equiv \frac{|x - x^*|}{|x|}$$

Para responder a pergunta, precisamos comparar a ordem de grandeza (o peso) dos **erros absolutos** com os respectivos valores exatos.

Vamos comparar por meio do **erro relativo**.

- O **erro relativo** de x^* como uma aproximação de x é defido por:

$$ER_x \equiv \frac{|x - x^*|}{|x|}$$

- O **erro relativo**[2] não está definido para o caso $x = 0$.

[2] Em alguns livros o erro é definido com o sinal oposto ao que é usado aqui. Tipicamente não faz quase diferença alguma a convenção que se utiliza, desde que o seja de forma consistente ao longo de todo o texto. Note que $x - x^*$ é a correção que deve ser adicionada ao valor de x^* para eliminar do erro. A "correção" e o "erro absoluto" têm a mesma magnitude, mas podem ter sinais diferentes.

Então temos que:

$$\frac{|30.1358 - 30.0000|}{30.1358} = 0.004506268$$
$$\frac{|1.1358 - 1|}{1.1358} = 0.119563303$$

Então temos que:

$$\frac{|30.1358 - 30.0000|}{30.1358} = 0.004506268$$
$$\frac{|1.1358 - 1|}{1.1358} = 0.119563303$$

Ou ainda, pelo **erro relativo percentual**,

$$\frac{|30.1358 - 30.0000|}{30.1358} = 0.004506268 \approx 0.45\%$$
$$\frac{|1.1358 - 1|}{1.1358} = 0.119563303 \approx 11.96\%$$

Então temos que:

$$\frac{|30.1358 - 30.0000|}{30.1358} = 0.004506268$$
$$\frac{|1.1358 - 1|}{1.1358} = 0.119563303$$

Ou ainda, pelo **erro relativo percentual**,

$$\frac{|30.1358 - 30.0000|}{30.1358} = 0.004506268 \approx 0.45\%$$
$$\frac{|1.1358 - 1|}{1.1358} = 0.119563303 \approx 11.96\%$$

Portanto, devemos utilizar o **erro relativo** para obter uma **melhor informação** sobre a precisão das aproximações e suas ordens de grandeza.

- **Sistema de Numeração Utilizado pelo Computador**

- **Fato**: O conjunto dos números representáveis **em qualquer** máquina é **finito**.

- **Fato**: O conjunto dos números representáveis **em qualquer** máquina é **finito**.

  Ou seja, **não** é possível representar em uma máquina todos os números de um dado intervalo $[a, b]$, $\quad a < b$.

- **OBS.:** Como vimos anteriormente, no exemplo da resolução de sistemas lineares, a implicação desse fato é que o resultado de uma **simples operação aritmética** ou o **cálculo de uma função**, realizadas com esses números, **podem conter erros**.

- **Representação de números em ponto flutuante NORMALIZADO**

  Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

  $$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e, \text{ onde}$$

- **Representação de números em ponto flutuante NORMALIZADO**

  Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

  $$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e, \text{ onde}$$

  β é a base de operações aritméticas da máquina

- **Representação de números em ponto flutuante NORMALIZADO**

Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e, \text{ onde}$$

β é a base de operações aritméticas da máquina

e é o **expoente**, $-m \leq e \leq M \quad (m, M \in \mathbb{N})$

- **Representação de números em ponto flutuante NORMALIZADO**

Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e, \text{ onde}$$

β é a base de operações aritméticas da máquina

e é o **expoente**, $-m \leq e \leq M \quad (m, M \in \mathbb{N})$

t é o número de dígitos da **mantissa**, $d_1 \neq 0$,

$0 \leq d_i < \beta$, $i = 1, 2, 3 \cdots t$.

- **Representação de números em ponto flutuante NORMALIZADO**

Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e, \text{ onde}$$

β é a base de operações aritméticas da máquina

e é o **expoente**, $-m \leq e \leq M \quad (m, M \in \mathbb{N})$

t é o número de dígitos da **mantissa**, $d_1 \neq 0$,

$0 \leq d_i < \beta$, $i = 1, 2, 3 \cdots t$.

$\rightarrow$ O número 0 (**zero**) pertence a qualquer sistema

Por simplicidade,

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e,\ d_1 \neq 0,\ -m \leq e \leq M,$$

será representado por $F(\beta, t, m, M)$

Por simplicidade,

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e,\ d_1 \neq 0,\ -m \leq e \leq M,$$

será representado por $F(\beta, t, m, M)$

EXEMPLO (considere o sistema $F(10, 3, 2, 2)$).

Por simplicidade,

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e,\ d_1 \neq 0,\ -m \leq e \leq M,$$

será representado por $F(\beta, t, m, M)$.

EXEMPLO (considere o sistema $F(10, 3, 2, 2)$). Ou seja, um número neste sistema será dado por:

$$\pm 0.d_1 d_2 d_3 \times 10^e,\ -2 \leq e \leq 2,$$
$$0 \leq d_i < 10,\ i = 1, 2, 3,\ d_1 \neq 0.$$

Por simplicidade,

$$\pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e,\ d_1 \neq 0,\ -m \leq e \leq M,$$

será representado por $F(\beta, t, m, M)$.

EXEMPLO (considere o sistema $F(10, 3, 2, 2)$). Ou seja, um número neste sistema será dado por:

$$\pm 0.d_1 d_2 d_3 \times 10^e,\ -2 \leq e \leq 2,$$
$$0 \leq d_i < 10,\ i = 1, 2, 3,\ d_1 \neq 0.$$

Vamos representar os números $+3.51$, -12.345678, -6.287, $+7.284$, -0.0003 e 5398.2, no sistema $F(10, 3, 2, 2)$.

- $+3.51 = +0.351 \times 10^1$ (exato).

- $+3.51 = +0.351 \times 10^1$ (exato).
- $-12.345678 = -0.123 \times 10^2$ (perda de dígitos signiticativos).

- $+3.51 = +0.351 \times 10^1$ (exato).
- $-12.345678 = -0.123 \times 10^2$ (perda de dígitos signiticativos).
- $-6.287 = -0.628 \times 10^1$ (perda de dígito signiticativo).

- $+3.51 = +0.351 \times 10^1$ (exato).
- $-12.345678 = -0.123 \times 10^2$ (perda de dígitos signiticativos).
- $-6.287 = -0.628 \times 10^1$ (perda de dígito signiticativo).
- $+7.284 = +0.728 \times 10^1$ (perda de dígito signiticativo).

- $+3.51 = +0.351 \times 10^1$ (exato).
- $-12.345678 = -0.123 \times 10^2$ (perda de dígitos signiticativos).
- $-6.287 = -0.628 \times 10^1$ (perda de dígito signiticativo).
- $+7.284 = +0.728 \times 10^1$ (perda de dígito signiticativo).
- $-0.0003 = -0.3 \times 10^{-3}$. Expoente $-3 < -2$.

  Neste caso temos um **underflow**, i.e., quando o resultado é muito pequeno para ser representado em um dado sistema.

- $+3.51 = +0.351 \times 10^1$ (exato).
- $-12.345678 = -0.123 \times 10^2$ (perda de dígitos signiticativos).
- $-6.287 = -0.628 \times 10^1$ (perda de dígito signiticativo).
- $+7.284 = +0.728 \times 10^1$ (perda de dígito signiticativo).
- $-0.0003 = -0.3 \times 10^{-3}$. Expoente $-3 < -2$.
  Neste caso temos um **underflow**, i.e., quando o resultado é muito pequeno para ser representado em um dado sistema.
- $+5398.2 = +0.539 \times 10^4$. Expoente $4 > 2$.
  Neste caso temos um **overflow**, i.e., quando o resultado é muito grande para ser representado em um dado sistema.

- Representação de números no sistema $F(\beta, t, m, M)$

- Representação de números no sistema $F(\beta, t, m, M)$

$\rightarrow$ Sabemos que os números reais podem ser representados por uma reta contínua.

$\rightarrow$ Entretanto, em **ponto flutuante** podemos representar apenas **pontos discretos** da reta real.

- Representação de números no sistema $F(\beta, t, m, M)$

$\rightarrow$ Sabemos que os números reais podem ser representados por uma reta contínua.

$\rightarrow$ Entretanto, em **ponto flutuante** podemos representar apenas **pontos discretos** da reta real.

Pergunta: **Quantos** e **quais** números podem ser representados no sistema $F(2, 3, 1, 2)$?

Solução: Para $F(2,3,1,2)$, temos que $\beta = 2$, então os dígitos podem ser 0 ou 1 ($0 \leq d_i < \beta$).

Além disso, $m = 1$ e $M = 2$, então $-1 \leq e \leq 2$ e $t = 3$, o número de dígitos significativos.

Assim, os números são da forma: $\pm 0.d_1 d_2 d_3 \times \beta^e$

Solução: Para $F(2,3,1,2)$, temos que $\beta = 2$, e que os **dígitos** d_i podem ser 0 ou 1 ($0 \leq d_i < \beta$).

Além disso, $m = 1$ e $M = 2$, então $-1 \leq e \leq 2$ e $t = 3$, o número de dígitos significativos.

Assim, os números são da forma: $\pm 0.d_1 d_2 d_3 \times \beta^e$

possibilidades para o sinal ($+$ ou $-$): **2**
possibilidades para d_1 ($d_1 \neq 0$): **1**
possibilidades para d_2 (0 ou 1): **2**
possibilidades para d_3 (0 ou 1): **2**
possibilidades para β^e ($\beta = 2$, $e = -1, 0, 1, 2$): **4**

Solução: Para $F(2,3,1,2)$, temos que $\beta = 2$, então os dígitos podem ser 0 ou 1 ($0 \leq d_i < \beta$).

Além disso, $m = 1$ e $M = 2$, então $-1 \leq e \leq 2$ e $t = 3$, o número de dígitos significativos.

Assim, os números são da forma: $\pm 0.d_1 d_2 d_3 \times \beta^e$

possibilidades para o sinal ($+$ ou $-$): **2**
possibilidades para d_1 ($d_1 \neq 0$): **1**
possibilidades para d_2 (0 ou 1): **2**
possibilidades para d_3 (0 ou 1): **2**
possibilidades para β^e ($\beta = 2$, $e = -1, 0, 1, 2$): **4**

Fazendo o produto: $2 \times 1 \times 2 \times 2 \times 4 =$ **32**

Como o 0 (zero) faz parte de qualquer sistema, podemos representar 33 números no sistema $F(2,3,1,2)$.

Como o 0 (zero) faz parte de qualquer sistema, podemos representar 33 números no sistema $F(2,3,1,2)$.

Quais são os números ?

Como o 0 (zero) faz parte de **qualquer sistema**, podemos representar 33 números no sistema $F(2, 3, 1, 2)$.

Quais são os números ?

As formas da mantissa: 0.100, 0.101, 0.110 e 0.111

As formas de β^e são: 2^{-1}, 2^0, 2^1 e 2^2

Como o 0 (zero) faz parte de **qualquer sistema**, podemos representar 33 números no sistema $F(2, 3, 1, 2)$.

Quais são os números ?

As formas da mantissa: 0.100, 0.101, 0.110 e 0.111

As formas de β^e são: 2^{-1}, 2^0, 2^1 e 2^2

Além do 0 (zero), obtemos então os seguintes números:

Como o 0 (zero) faz parte de qualquer sistema, podemos representar 33 números no sistema $F(2,3,1,2)$.

Quais são os números ?

As formas da mantissa: 0.100, 0.101, 0.110 e 0.111
As formas de β^e são: 2^{-1}, 2^0, 2^1 e 2^2

Além do 0 (zero), obtemos então os seguintes números:

$\times$	2^{-1}	2^0	2^1	2^2
± 0.100	± 0.25	± 0.5	± 1.0	± 2.0
± 0.101	± 0.3125	± 0.625	± 1.25	± 2.5
± 0.110	± 0.375	± 0.750	± 1.5	± 3.0
± 0.111	± 0.4375	± 0.875	± 1.75	± 3.5

Como o 0 (zero) faz parte de qualquer sistema, podemos representar 33 números no sistema $F(2,3,1,2)$.

As formas da mantissa: 0.100, 0.101, 0.110 e 0.111
As formas de β^e são: 2^{-1}, 2^0, 2^1 e 2^2

Além do 0 (zero), obtemos então os seguintes números:

$\times$	2^{-1}	2^0	2^1	2^2
± 0.100	± 0.25	± 0.5	± 1.0	± 2.0
± 0.101	± 0.3125	± 0.625	± 1.25	± 2.5
± 0.110	± 0.375	± 0.750	± 1.5	± 3.0
± 0.111	± 0.4375	± 0.875	± 1.75	± 3.5

- **Fato**: Com base nos modelos existentes (viáveis) de aritmética de ponto flutuante sempre teremos para quaisquer t, m e M, um sistema finito de números !!!

- Para exemplificar um pouco mais as limitações encontradas nos computadores, considere o seguinte exemplo.

- Para exemplificar um pouco mais as limitações encontradas nos computadores, considere o seguinte exemplo.

  EXEMPLO. Seja $f(x)$ uma função contínua real, definida no intervalo $[a, b]$, $a < b$. Sejam $f(a) < 0$ e $f(b) > 0$.

- Para exemplificar um pouco mais as limitações encontradas nos computadores, considere o seguinte exemplo.

  EXEMPLO. Seja $f(x)$ uma função contínua real, definida no intervalo $[a, b]$, $a < b$. Sejam $f(a) < 0$ e $f(b) > 0$.

  Então de acordo com o teorema do valor intermediário, existe $x \in [a, b]$ tal que $f(x) = 0$.

- Para exemplificar um pouco mais as limitações encontradas nos computadores, considere o seguinte exemplo.

  EXEMPLO. Seja $f(x)$ uma função contínua real, definida no intervalo $[a, b]$, $a < b$. Sejam $f(a) < 0$ e $f(b) > 0$.

  Então de acordo com o teorema do valor intermediário, existe $x \in [a, b]$ tal que $f(x) = 0$.

  Seja $f(x) = x^3 - 3 \quad , x \in [a_k, b_k] \subset [a, b]$, $k = 1, 2, 3, \cdots$

- Para exemplificar um pouco mais as limitações encontradas nos computadores, considere o seguinte exemplo.

EXEMPLO. Seja $f(x)$ uma função contínua real, definida no intervalo $[a, b]$, $a < b$. Sejam $f(a) < 0$ e $f(b) > 0$.

Então de acordo com o teorema do valor intermediário, existe $x \in [a, b]$ tal que $f(x) = 0$.

Seja $f(x) = x^3 - 3 \quad , x \in [a_k, b_k] \subset [a, b]$, $k = 1, 2, 3, \cdots$

Vamos determinar x tal que $f(x) = 0$ para um sistema $F(10, 10, 10, 10)$.

Para a função $f(x) = x^3 - 3$, pode-se obter os resultados:
(sendo a_k e b_k obtidos por um método de aproximação)

$f(a_k = 0.1442249570 \times 10^1) = -0.2 \times 10^{-8}$.
$f(b_k = 0.1442249571 \times 10^1) = +0.4 \times 10^{-8}$.

Para a função $f(x) = x^3 - 3$, pode-se obter os resultados:
(sendo a_k e b_k obtidos por um método de aproximação)

$f(a_k = 0.1442249570 \times 10^1) = -0.2 \times 10^{-8}$.
$f(b_k = 0.1442249571 \times 10^1) = +0.4 \times 10^{-8}$.

Observe que entre a_k e b_k ($a_k < b_k$),

$$a_k = 0.1442249570 \times 10^1 \text{ e } b_k = 0.1442249571 \times 10^1$$

não existe algum número que possa ser representado no **sistema dado** $F(10, 10, 10, 10)$, e que a função $f(x)$ muda de sinal nos extremos desse intervalo.

Para a função $f(x) = x^3 - 3$, pode-se obter os resultados: (sendo a_k e b_k obtidos por um método de aproximação)

$f(a_k = 0.1442249570 \times 10^1) = -0.2 \times 10^{-8}$.
$f(b_k = 0.1442249571 \times 10^1) = +0.4 \times 10^{-8}$.

Observe que entre a_k e b_k ($a_k < b_k$),
$a_k = 0.1442249570 \times 10^1$ e $b_k = 0.1442249571 \times 10^1$
não existe algum número que possa ser representado no **sistema dado** $F(10, 10, 10, 10)$, e que a função $f(x)$ muda de sinal nos extremos desse intervalo.

- Assim, nesta "máquina", não é possível representar um número $x \in [a_k, b_k] \subset [a, b]$ tal que $f(x) = 0$. Portanto, a equação $f(x) = x^3 - 3 = 0$, **não** possui solução no sistema $F(10, 10, 10, 10)$!

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

  Conforme vimos anteriormente, é importante quantificar o erro que se comente em cálculos numéricos

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

  Conforme vimos anteriormente, é importante quantificar o erro que se comente em cálculos numéricos

  E a representação de um número para cálculos numéricos depende intrinsecamente das características de cada máquina (e.g., base β, mantissa t, precisão *simples* ou *dupla*, $\cdots$)

Erros de Arredondamento e Truncamento em Ponto Flutuante

Conforme vimos anteriormente, é importante quantificar o erro que se comente em cálculos numéricos

E a representação de um número para cálculos numéricos depende intrinsecamente das características de cada máquina (e.g., base β, mantissa t, precisão *simples* ou *dupla*, $\cdots$)

Considere uma máquina que opera em aritmética de ponto flutuante, com t dígitos significativos e base 10.

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$
$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$
$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$
$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$
$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}$$

EXEMPLO: $z = 98765.4321$ e $t = 6$ dígitos na mantissa

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$
$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$
$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}$$

EXEMPLO: $z = 98765.4321$ e $t = 6$ dígitos na mantissa

$$z = 0.987654321 \times 10^5$$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):
$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa
$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$
$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa
$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$
$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}$

EXEMPLO: $z = 98765.4321$ e $t = 6$ dígitos na mantissa
$z = 0.987654321 \times 10^5$
$z = 0.987654 \times 10^5 + 0.000000321 \times 10^5$

Nesta máquina, um número x pode ser representado da seguinte forma (conveniente):
$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa
$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$
$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa
$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$
$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}$

EXEMPLO: $z = 98765.4321$ e $t = 6$ dígitos na mantissa
$z = 0.987654321 \times 10^5$
$z = 0.987654 \times 10^5 + 0.000000321 \times 10^5$
$z = 0.987654 \times 10^5 + 0.321 \times 10^{(5-6)=-1}$

- **Lembrando...**

  Queremos estudar os erros (**absoluto** e **relativo**) nos processos de arredondamento e truncamento em aritimética de ponto flutuante

- **Lembrando...**

  Queremos estudar os erros (**absoluto** e **relativo**) nos processos de arredondamento e truncamento em aritimética de ponto flutuante

  Em uma máquina, o número

  $x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$

  pode ser representado por truncamento ou arredondamento.

- **Lembrando...**

  Queremos estudar os erros (**absoluto** e **relativo**) nos processos de arredondamento e truncamento em aritimética de ponto flutuante

  Em uma máquina, o número
  $x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$
  pode ser representado por truncamento ou arredondamento.

- **Truncamento** ($\overline{x}$ poder ser visto como x^*)

  A quantidade $g_x \times 10^{e-t}$ é descartada e $\overline{x} = f_x \times 10^e$

- **Lembrando...**

  Queremos estudar os erros (**absoluto** e **relativo**) nos processos de arredondamento e truncamento em aritimética de ponto flutuante

  Em uma máquina, o número
  $x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$
  pode ser representado por truncamento ou arredondamento.

- **Truncamento** ($\overline{x}$ poder ser visto como x^*)
  A quantidade $g_x \times 10^{e-t}$ é descartada e $\overline{x} = f_x \times 10^e$

  Temos ainda que:

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$$

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$$
$$EA_x = |g_x \times 10^{e-t}|$$

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x \times 10^{e-t}|$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$$

$$EA_x = |g_x \times 10^{e-t}|$$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$$

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x \times 10^{e-t}|$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} \text{ (Erro Relativo)}$$

Erros de Arredondamento e Truncamento em Ponto Flutuante

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$$

$$EA_x = |g_x \times 10^{e-t}|$$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e}$$

Erros de Arredondamento e Truncamento em Ponto Flutuante

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$$

$$EA_x = |g_x \times 10^{e-t}|$$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{|f_x| \times 10^e}$$

Erros de Arredondamento e Truncamento em Ponto Flutuante

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x \times 10^{e-t}|$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{0.1 \times 10^e}$$

Erros de Arredondamento e Truncamento em Ponto Flutuante

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$$

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x \times 10^{e-t}|$

Como $0 \leq g_x < 1$, temos que $|g_x| < 1$. E segue que:

$EA_x = |g_x \times 10^{e-t}| < 10^{e-t}$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{0.1 \times 10^e} = 10^{1-t}$$

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

  $x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$

- **Arredondamento**

  f_x é alterado para levar em conta a quantidade g_x

- **Erros de Arredondamento e Truncamento em Ponto Flutuante**

  $x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1,$

- **Arredondamento**

  f_x é alterado para levar em conta a quantidade g_x

  Uma forma de arredondamento amplamente empregada é o *arredondamento simétrico*, dado por:

$$\overline{x} = \begin{cases} f_x \times 10^e & \text{se} \quad |g_x| < \dfrac{1}{2} \\ f_x \times 10^e + 10^{e-t} & \text{se} \quad |g_x| \geq \dfrac{1}{2} \end{cases}$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}, \; g_x = 0.7 \geq \frac{1}{2}$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}, \; g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1}$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}, \; g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1},\ g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}, \; g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1}, \; g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1}, \; g_x = 0.123 < \frac{1}{2}$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1},\ g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1},\ g_x = 0.123 < \frac{1}{2}$$

$$y = 0.789 \times 10^4$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1},\ g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1},\ g_x = 0.123 < \frac{1}{2}$$

$$y = 0.789 \times 10^4 = 7890$$

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \quad 0.1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1.$$

EXEMPLO: $x = 234.57$ e $t = 4$ dígitos na mantissa

$$x = 0.23457 \times 10^3 = 0.2345 \times 10^3 + 0.00007 \times 10^3$$

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{(3-4)=-1},\ g_x = 0.7 \geq \frac{1}{2}$$

$$x = 0.2345 \times 10^3 + 10^{-1} = 234.5 + 0.1 = 234.6$$

EXEMPLO: $y = 7891.23$ e $t = 3$ dígitos na mantissa

$$y = 0.789123 \times 10^4 = 0.789 \times 10^4 + 0.000123 \times 10^4$$

$$y = 0.789 \times 10^4 + 0.123 \times 10^{(4-3)=1},\ g_x = 0.123 < \frac{1}{2}$$

$$y = 0.789 \times 10^4 = 7890$$

- Vamos analisar EA_x e ER_x para $|g_x| < \frac{1}{2}$ e $|g_x| \geq \frac{1}{2}$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

  $EA_x = |g_x| \times 10^{e-t}$.

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

  $EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

  $EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

  $EA_x < \frac{1}{2} \times 10^{e-t}$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{0.1 \times 10^e}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| < \frac{1}{2}$, temos:

$EA_x = |x - \overline{x}|$ (Erro Absoluto)

$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e)|$

$EA_x = |g_x| \times 10^{e-t}$. Como $|g_x| < \frac{1}{2}$, segue:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{0.1 \times 10^e}$$

$$ER_x < \frac{1}{2} \times 10^{1-t}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$

  $EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

  $EA_x = |x - \overline{x}|$ (Erro Absoluto)

  $EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$

  $EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$

  $EA_x = |g_x - 1| \times 10^{e-t}$ (note também que $0 \leq g_x < 1$),

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \ (0 \leq g_x < 1),$$
$$EA_x \leq \frac{1}{2} \times 10^{e-t}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \; (0 \leq g_x < 1),$$
$$EA_x \leq \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \; (0 \leq g_x < 1),$$
$$EA_x \leq \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$
$$ER_x < \frac{0.5 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \; (0 \leq g_x < 1),$$
$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$
$$ER_x < \frac{0.5 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \; (0 \leq g_x < 1),$$
$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$

$$ER_x < \frac{0.5 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{0.1 \times 10^e}$$

- **Erros de Arredondamento e Truncamento**
- Se $|g_x| \geq \frac{1}{2}$, temos:

$$EA_x = |x - \overline{x}| \text{ (Erro Absoluto)}$$
$$EA_x = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})|$$
$$EA_x = |g_x \times 10^{e-t} - 10^{e-t}|$$
$$EA_x = |g_x - 1| \times 10^{e-t} \; (0 \leq g_x < 1),$$
$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x = \frac{|x - \overline{x}|}{|\overline{x}|} = \frac{EA_x}{|\overline{x}|} \text{ (Erro Relativo)}$$
$$ER_x < \frac{0.5 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} < \frac{0.5 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{0.1 \times 10^e}$$
$$ER_x < \frac{1}{2} \times 10^{1-t}$$

- **Erros de Arredondamento e Truncamento**

- Assim, para ambos os casos, os erros são dados por:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x < \frac{1}{2} \times 10^{1-t}$$

- **Erros de Arredondamento e Truncamento**

- Assim, para ambos os casos, os erros são dados por:

$$EA_x < \frac{1}{2} \times 10^{e-t}$$

$$ER_x < \frac{1}{2} \times 10^{1-t}$$

- O erro cometido no processo de arredondamento é menor do que aquele apresentado no truncamento.

- **Erros de Arredondamento e Truncamento**

- A análise revela que os erros são dados por:

$$EA_x < \frac{1}{2} \times 10^{e-t} \qquad (\text{“Se” } t \to \infty \text{ “então” } EA_x \to 0)$$

$$ER_x < \frac{1}{2} \times 10^{1-t} \qquad (\text{“Se” } t \to \infty \text{ “então” } ER_x \to 0)$$

- O erro cometido no processo de arredondamento é menor do que aquele apresentado no truncamento.

- Por outro lado, do ponto de vista de esforço computacional (tempo de execução das operações), o truncamento requer menos tempo do que o arredondamento.

- **Operações Aritméticas em Ponto Flutuante**

- Considere uma máquina e/ou computador qualquer e uma série de **operações aritméticas**.

- Considere uma máquina e/ou computador qualquer e uma série de **operações aritméticas**.

  Pelo fato do **arredondamento/truncamento** ser feito após **cada operação** temos, ao contrário do que é **válido para números reais**, que as operações aritméticas

  adição, **subtração**, **divisão** e **multiplicação**

  não são **nem associativas** e **nem distributivas**.

- Considere uma máquina e/ou computador qualquer e uma série de **operações aritméticas**.

  Pelo fato do **arredondamento**/**truncamento** ser feito após **cada operação** temos, ao contrário do que é **válido para números reais**, que as operações aritméticas

  adição, **subtração**, **divisão** e **multiplicação**

  não são **nem associativas** e **nem distributivas**.

  Vejamos esse fato por meio de exemplos.

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas:

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

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas usando o **arredondamento simétrico**:

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

$$(11.4 + 3.18) + 5.05 = 14.6 + 5.05 = 19.7$$

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas usando o **arredondamento simétrico**:

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

$$(11.4 + 3.18) + 5.05 = 14.6 + 5.05 = 19.7$$
$$11.4 + (3.18 + 5.05) = 11.4 + 8.23 = 19.6$$

- Calcule as expressões indicadas nos itens b) e c):

b) $\dfrac{3.18 \times 11.4}{5.05}$ e $\dfrac{3.18}{5.05} \times 11.4$

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

b) $\dfrac{3.18 \times 11.4}{5.05}$ e $\dfrac{3.18}{5.05} \times 11.4$

$$\frac{3.18 \times 11.4}{5.05} = \frac{36.3}{5.05} = 7.19$$

$$\frac{3.18}{5.05} \times 11.4 = 0.630 \times 11.4 = 7.18$$

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

$$3.18 \times (5.05 + 11.4) = 3.18 \times 16.5 = 52.5$$

$$3.18 \times 5.05 + 3.18 \times 11.4 = 16.1 + 36.3 = 52.4$$

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Valor exato: considere **todos** os dígitos de uma máquina, sem usar arredondamento a cada operação.

Segue que, $P(5.24) = -0.00776$ (**valor exato**).

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Valor exato: considere **todos** os dígitos de uma máquina, sem usar arredondamento a cada operação.

Segue que, $P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.

$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$

$= 144. - 165. + 21.0 - 0.1$

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Valor exato: considere **todos** os dígitos de uma máquina, sem usar arredondamento a cada operação.

Segue que, $P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.

$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$

$= 144. - 165. + 21.0 - 0.1$

$= -0.10$ (somando da esquerda para a direita)

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Valor exato: considere **todos** os dígitos de uma máquina, sem usar arredondamento a cada operação.

Segue que, $P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.

$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$

$= 144. - 165. + 21.0 - 0.1$

$= -0.10$ (somando da esquerda para a direita)

$= 0.00$ (somando da direita para a esquerda).

Continuando. Note ainda que o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x(x(x - 6) + 4) - 0.1$$

(forma computacional “tipicamente” eficiente)

Continuando. Note ainda que o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x(x(x-6)+4) - 0.1$$

Assim:

$P(5.24) = 5.24(5.24(5.24-6)+4) - 0.1$

$= 5.24(-3.98+4) - 0.1$

$= 5.24(0.02) - 0.1$

$= 0.105 - 0.1$

$= 0.005$ (sinal errado !!!)

Continuando. Note ainda que o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x(x(x-6)+4) - 0.1$$

Assim:

$$P(5.24) = 5.24(5.24(5.24-6)+4) - 0.1$$
$$= 5.24(-3.98+4) - 0.1$$
$$= 5.24(0.02) - 0.1$$
$$= 0.105 - 0.1$$
$$= 0.005 \text{ (sinal errado !!!)}$$

- A perda de dígitos significativos requer cuidados para evitar impacto negativo na computação em precisão finita.

- **Análise de Erros nas Operações Aritméricas de ponto Flutuante**

  No que segue indicam-se as fórmulas para os erros absoluto e relativo nas operações aritméticas em ponto flutuante

- **Análise de Erros nas Operações Aritméricas de ponto Flutuante**

  No que segue indicam-se as fórmulas para os erros absoluto e relativo nas operações aritméticas em ponto flutuante

  Nas contas que seguem supõe-se que o erro final é arredondado.

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x + y$

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

  Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

  Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

  Erro Relativo

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

  Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

  Erro Relativo

  $$ER_{x+y} = \frac{EA_{x+y}}{\overline{x} + \overline{y}}$$

- **Adição**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

  Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

  Erro Relativo

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

Adição

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

Erro Relativo

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

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

Adição

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

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

Daí segue que o erro absoluto EA_{x+y} da soma vale:

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

Erro Relativo

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

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

$$ER_{x+y} = ER_x \left(\frac{\overline{x}}{\overline{x} + \overline{y}} \right) + ER_y \left(\frac{\overline{y}}{\overline{x} + \overline{y}} \right)$$

- **Subtração**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

- Analogamente ao caso da adição, temos:

- **Subtração**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

- Analogamente ao caso da adição, temos:

  Erro Absoluto

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

- **Subtração**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

- Analogamente ao caso da adição, temos:

  Erro Absoluto

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

  Erro Relativo

  $$ER_{x+y} = ER_x \left(\frac{\overline{x}}{\overline{x} - \overline{y}} \right) - ER_y \left(\frac{\overline{y}}{\overline{x} - \overline{y}} \right)$$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\ y$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$$

  $$x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$

  $x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$

  $x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

  $EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$

  $x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

  $EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$

  Erro Relativo

Multiplicação

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$$

$$x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$$

Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

$$EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$$

Erro Relativo

$$ER_{x\,y} \approx \frac{EA_{x\,y}}{\overline{x}\;\overline{y}}$$

- **Multiplicação**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$

  $x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

  $EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$

  Erro Relativo

  $$ER_{x\,y} \approx \frac{EA_{x\,y}}{\overline{x}\;\overline{y}} = \frac{\overline{x}EA_y + \overline{y}EA_x}{\overline{x}\;\overline{y}}$$

Multiplicação

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$$

$$x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$$

Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

$$EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$$

Erro Relativo

$$ER_{x\,y} \approx \frac{EA_{x\,y}}{\overline{x}\;\overline{y}} = \frac{\overline{x}EA_y + \overline{y}EA_x}{\overline{x}\;\overline{y}} = \frac{EA_x}{\overline{x}} + \frac{EA_y}{\overline{y}}$$

Multiplicação

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$x\,y = (\overline{x} + EA_x)\,(\overline{y} + EA_y)$$

$$x\,y = \overline{x}\;\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_x\,EA_y$$

Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

$$EA_{x\,y} \approx \overline{x}EA_y + \overline{y}EA_x$$

Erro Relativo

$$ER_{x\,y} \approx \frac{EA_{x\,y}}{\overline{x}\;\overline{y}} = \frac{\overline{x}EA_y + \overline{y}EA_x}{\overline{x}\;\overline{y}} = \frac{EA_x}{\overline{x}} + \frac{EA_y}{\overline{y}}$$

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

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $\frac{x}{y}$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y}$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y} = \frac{\overline{x} + EA_x}{\overline{y}} \left(\frac{1}{1 + \frac{EA_y}{\overline{y}}} \right)$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y} = \frac{\overline{x} + EA_x}{\overline{y}} \left(\frac{1}{1 + \frac{EA_y}{\overline{y}}} \right)$$

  Atenção!

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y} = \frac{\overline{x} + EA_x}{\overline{y}} \left(\frac{1}{1 + \frac{EA_y}{\overline{y}}} \right)$$

  Atenção!

  Por conveniência na análise, vamos expressar o termo $\frac{1}{1 + \frac{EA_y}{\overline{y}}}$ como segue:

Divisão

Erro Absoluto

Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y} = \frac{\overline{x} + EA_x}{\overline{y}} \left(\frac{1}{1 + \frac{EA_y}{\overline{y}}} \right)$$

Atenção!

Por conveniência na análise, vamos expressar o termo $\frac{1}{1 + \frac{EA_y}{\overline{y}}}$ como segue (série infinita):

$$\frac{1}{1 + \frac{EA_y}{\overline{y}}} = 1 - \frac{EA_y}{\overline{y}} + \left(\frac{EA_y}{\overline{y}} \right)^2 - \left(\frac{EA_y}{\overline{y}} \right)^3 + \cdots$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} = \frac{\overline{x} + EA_x}{\overline{y} + EA_y} = \frac{\overline{x} + EA_x}{\overline{y}} \left(\frac{1}{1 + \frac{EA_y}{\overline{y}}} \right)$$

  Atenção!

  Por conveniência na análise, vamos expressar o termo $\frac{1}{1 + \frac{EA_y}{\overline{y}}}$ como segue (série infinita):

  $$\frac{1}{1 + \frac{EA_y}{\overline{y}}} = 1 - \frac{EA_y}{\overline{y}} + \left(\frac{EA_y}{\overline{y}} \right)^2 - \left(\frac{EA_y}{\overline{y}} \right)^3 + \cdots$$

  e desprezando os termos com potências maiores do que 1, temos:

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right)$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right) = \left(\frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}}\right) \left(1 - \frac{EA_y}{\overline{y}}\right)$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

$$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right) = \left(\frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}}\right) \left(1 - \frac{EA_y}{\overline{y}}\right)$$

$$\frac{x}{y} \approx \frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} - \frac{EA_x\, EA_y}{\overline{y}^2}$$

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right) = \left(\frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}}\right) \left(1 - \frac{EA_y}{\overline{y}}\right)$$

  $$\frac{x}{y} \approx \frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} - \frac{EA_x\, EA_y}{\overline{y}^2}$$

  Atenção!

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right) = \left(\frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}}\right) \left(1 - \frac{EA_y}{\overline{y}}\right)$$

  $$\frac{x}{y} \approx \frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} - \frac{EA_x\, EA_y}{\overline{y}^2}$$

  Atenção!

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

- **Divisão**

  Erro Absoluto

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  $$\frac{x}{y} \approx \frac{\overline{x} + EA_x}{\overline{y}} \left(1 - \frac{EA_y}{\overline{y}}\right) = \left(\frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}}\right) \left(1 - \frac{EA_y}{\overline{y}}\right)$$

  $$\frac{x}{y} \approx \frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} - \frac{EA_x\, EA_y}{\overline{y}^2}$$

  Atenção!

  Supondo que o produto $(EA_x)\,(EA_y)$ é um número pequeno, decarta-se então este termo da última equação.

  Segue que:

  $$\frac{x}{y} \approx \frac{\overline{x}}{\overline{y}} + \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

$$ER_{x/y} \approx \frac{EA_{x/y}}{\frac{x}{y}}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

$$ER_{x/y} \approx \frac{EA_{x/y}}{\frac{\overline{x}}{\overline{y}}} = \left(\frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2} \right) \frac{\overline{y}}{\overline{x}}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

$$ER_{x/y} \approx \frac{EA_{x/y}}{\frac{\overline{x}}{\overline{y}}} = \left(\frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2} \right) \frac{\overline{y}}{\overline{x}}$$

$$ER_{x/y} \approx \frac{EA_x}{\overline{x}} - \frac{EA_y}{\overline{y}}$$

- **Divisão**

  Erro Absoluto

  Finalmente,

$$EA_{x/y} = \frac{EA_x}{\overline{y}} - \frac{\overline{x}\, EA_y}{\overline{y}^2} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

$$ER_{x/y} \approx \frac{EA_{x/y}}{\frac{\overline{x}}{\overline{y}}} = \left(\frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2} \right) \frac{\overline{y}}{\overline{x}}$$

$$ER_{x/y} \approx \frac{EA_x}{\overline{x}} - \frac{EA_y}{\overline{y}} = ER_x - ER_y$$

- É válido lembrar que em todas as fórmulas, não foi considerado o erro de **arredondamento** ou **truncamento** no resultado final.

  Para uma análise completa, detalhada, da propagação dos erros, deve-se considerar ainda os erros em cada operação efetuada.

– Condicionamento de algoritmos & Efeitos Numéricos

Neste curso vamos examinar alguns procedimentos de aproximação, chamados **algoritmos**, que envolve uma sequência de cálculos.

Neste curso vamos examinar alguns procedimentos de aproximação, chamados **algoritmos**, que envolve uma sequência de cálculos.

Um **algoritmo** é um procedimento que descreve, de forma inequívoca, uma sequência finita de passos a ser realizada em uma ordem específica.

Neste curso vamos examinar alguns procedimentos de aproximação, chamados **algoritmos**, que envolve uma sequência de cálculos.

Um **algoritmo** é um procedimento que descreve, de forma inequívoca, uma sequência finita de passos a ser realizada em uma ordem específica.

O objetivo de um algoritmo é o de implementar um procedimento para a resolver um problema **ou** para uma solução aproximada do problema.

Além dos problemas dos erros causados pelas **operações aritméticas** existem certos **efeitos numéricos** que contribuem para que um resultado numérico **não seja confiável**.

Além dos problemas dos erros causados pelas **operações aritméticas** existem certos **efeitos numéricos** que contribuem para que um resultado numérico **não seja confiável**.

Alguns dos problemas mais frequentes em cálculo numérico são:

- Cancelamento
- Geração e Propagação de Erros
- Instabilidade Numérica
- Mal Condicionamento

Além dos problemas dos erros causados pelas **operações aritméticas** existem certos **efeitos numéricos** que contribuem para que um resultado numérico **não seja confiável**.

Alguns dos problemas mais frequentes em cálculo numéricos são:

- Cancelamento
- Geração e Propagação de Erros
- Instabilidade Numérica
- Mal Condicionamento

Vejamos alguns exemplos para ilustrar esse problemas

- **Cancelamento**

- **Cancelamento**

  O cancelamento ocorre tipicamente na **subtração** de dois números “quase iguais”, ou na adição de dois números com sinais opostos e “quase iguais” em **valor absoluto**

- **Cancelamento**

  O cancelamento ocorre tipicamente na **subtração** de dois números “quase iguais”, ou na adição de dois números com sinais opostos e “quase iguais” em **valor absoluto**

  Importante: Lembrem-se que estamos operando com aritmética de ponto flutuante !!

Cancelamento

O cancelamento ocorre tipicamente na **subtração** de dois números “quase iguais”, ou na adição de dois números com sinais opostos e “quase iguais” em **valor absoluto**

Importante: Lembrem-se que estamos operando com aritmética de ponto flutuante !!

Sem perda de generalidade, vamos supor que estamos trabalhando com o sistema $F(10, 10, 10, 10)$.

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$.

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$. Segue que:

$$\sqrt{9876} = 0.9937806599 \times 10^2 \text{ e}$$

$$\sqrt{9875} = 0.9937303457 \times 10^2$$

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$. Segue que:

$$\sqrt{9876} = 0.9937806599 \times 10^2 \text{ e}$$

$$\sqrt{9875} = 0.9937303457 \times 10^2$$

Então $\sqrt{9876} - \sqrt{9875} = 0.0000503142 \times 10^2$

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$. Segue que:

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

Então $\sqrt{9876} - \sqrt{9875} = 0.0000503142 \times 10^2$

A normalização muda este resultado para:

$$0.5031420000 \times 10^{-4}$$

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$. Segue que:

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

Então $\sqrt{9876} - \sqrt{9875} = 0.0000503142 \times 10^2$

A normalização muda este resultado para:

$$0.5031420000 \times 10^{-4}$$

Assim os quatro zeros no final da mantissa não têm significado e assim "**perdem-se**" 4 casas decimais.

EXEMPLO: $\sqrt{9876} - \sqrt{9875}$. Segue que:

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

Então $\sqrt{9876} - \sqrt{9875} = 0.0000503142 \times 10^2$

A normalização muda este resultado para:

$$0.5031420000 \times 10^{-4}$$

Assim os quatro zeros no final da mantissa não têm significado e assim "**perdem-se**" 4 casas decimais.

Pergunta: É possível obter um resultado mais preciso ?

Para este caso a resposta é **sim**.

Para este caso a resposta é **sim**. Basta consider a identidade:

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}.$$

Para este caso a resposta é **sim**. Basta consider a identidade:

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}.$$

Então, $\sqrt{9876} - \sqrt{9875} =$

$$0.5031418679 \times 10^{-4}$$

$$0.5031420000 \times 10^{-4} \text{ (antes)}$$

Para este caso a resposta é **sim**. Basta consider a identidade:

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}.$$

Então, $\sqrt{9876} - \sqrt{9875} =$

$$0.5031418679 \times 10^{-4}$$

$$0.5031420000 \times 10^{-4} \text{ (antes)}$$

Em **algumas** situações é possível explorar propriedades especiais de funções.

Por exemplo, se x e y são números "quase iguais", é conveniente substituir:

- $\sqrt{x} - \sqrt{y}$ por $\dfrac{x - y}{\sqrt{x} + \sqrt{y}}$
- $cos^2\,\theta - sen^2\theta$ por $cos\,(2\theta)$
- $log\,y - log\,x$ por $log\,\dfrac{y}{x}$
- $sen\,y - sen\,x$ por $2\,sen\,\frac{1}{2}(x - y)\,cos\,\frac{1}{2}(x + y)$

- Nos exemplos discutidos foi **possível** identificar e resolver o problema do cancelamento.

  Entretanto, é válido observar que **nem sempre** será possível indentificar uma maneira trivial de resolver problemas ocasionados pelo cancelamento.

- **Geração e Propagação de Erros**

Atenção !

Atenção !

Vamos revisitivar, e ampliar, a discussão de alguns exemplos **já apresentados**

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas:

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

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas:

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

$$(11.4 + 3.18) + 5.05 = 14.6 + 5.05 = 19.7$$

Sem perda de generalidade, considere o sistema com base $\beta = 10$ e 3 dígitos significativos.

Vamos efetuar os cálulos das expressões numéricas indicadas:

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

$$(11.4 + 3.18) + 5.05 = 14.6 + 5.05 = 19.7$$
$$11.4 + (3.18 + 5.05) = 11.4 + 8.23 = 19.6$$

- **Geração de erro na adição**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  Erro Absoluto

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

  Erro Relativo

  $$ER_{x+y} = ER_x \left(\frac{\overline{x}}{\overline{x} + \overline{y}} \right) + ER_y \left(\frac{\overline{y}}{\overline{x} + \overline{y}} \right)$$

- **Geração de erro na subtração**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  Erro Absoluto

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

  Erro Relativo

  $$ER_{x+y} = ER_x \left(\frac{\overline{x}}{\overline{x} - \overline{y}} \right) - ER_y \left(\frac{\overline{y}}{\overline{x} - \overline{y}} \right)$$

b) $\dfrac{3.18 \times 11.4}{5.05}$ e $\dfrac{3.18}{5.05} \times 11.4$

$$\frac{3.18 \times 11.4}{5.05} = \frac{36.3}{5.05} = 7.19$$

$$\frac{3.18}{5.05} \times 11.4 = 0.630 \times 11.4 = 7.18$$

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

$$3.18 \times (5.05 + 11.4) = 3.18 \times 16.5 = 52.5$$

$$3.18 \times 5.05 + 3.18 \times 11.4 = 16.1 + 36.3 = 52.4$$

- **Geração de erro na multiplicação**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  Erro Absoluto

  $EA_{x\,y} \approx \overline{x} EA_y + \overline{y} EA_x$

  Erro Relativo

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

- **Geração de erro na divisão**

  Sejam x e y, tais que $x = \overline{x} + EA_x$ e $y = \overline{y} + EA_y$.

  Erro Absoluto

  $$EA_{x/y} = \frac{\overline{y}\, EA_x - \overline{x}\, EA_y}{\overline{y}^2}$$

  Erro Relativo

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

Atenção !

Atenção !

Vamos ver como a **propagação de erro** ocorre $\cdots$

Atenção !

Vamos ver como a **propagação de erro** ocorre $\cdots$

em operações aritiméticas em ponto flutuante

Atenção !

Vamos ver como a **propagação de erro** ocorre $\cdots$

em operações aritiméticas em ponto flutuante

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Atenção !

Vamos ver como a **propagação de erro** ocorre $\cdots$

em operações aritiméticas em ponto flutuante

d) Calcular o polinômio $P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$ e comparar com o resultado exato.

Valor exato: considere **todos** os dígitos de uma máquina, sem usar arredondamento a cada operação.

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,

$P(5.24) = -0.00776$ (**valor exato**).

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,
$P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.
$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$
$= 144. - 165. + 21.0 - 0.1$

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,
$P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.
$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$
$= 144. - 165. + 21.0 - 0.1$
$= -0.10$ (somando da esquerda para a direita)

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,
$P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.
$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$
$= 144. - 165. + 21.0 - 0.1$
$= -0.10$ (somando da esquerda para a direita)
$= 0.00$ (somando da direita para a esquerda).

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,
$P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.
$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$
$= 144. - 165. + 21.0 - 0.1$
$= -0.10$ (somando da esquerda para a direita)
$= 0.00$ (somando da direita para a esquerda).

- Note que o erro cometido em cada parcela é propagado para as operações posteriores.

$P(x) = x^3 - 6x^2 + 4x - 0.1$ no ponto $x = 5.24$,
$P(5.24) = -0.00776$ (**valor exato**).

Usando arredondamento a cada operação efetuada.
$P(5.24) = 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1$
$= 144. - 165. + 21.0 - 0.1$
$= -0.10$ (somando da esquerda para a direita)
$= 0.00$ (somando da direita para a esquerda).

- Note que o erro cometido em cada parcela é propagado para as operações posteriores.

- Note também como a solução final difere do valor correto !

$P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x \cdot x \cdot x - 6 \cdot x \cdot x + 4 \cdot x - 0.1 \qquad (\mathbf{1})$$
$$P(x) = x(x(x - 6) + 4) - 0.1 \qquad (\mathbf{2})$$

$P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x \cdot x \cdot x - 6 \cdot x \cdot x + 4 \cdot x - 0.1 \qquad (\mathbf{1})$$
$$P(x) = x(x(x - 6) + 4) - 0.1 \qquad (\mathbf{2})$$

Em (1), 5 multiplicações, 2 subtrações e 1 adição

Em (2), 3 multiplicações, 2 subtrações e 1 adição

$P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x \cdot x \cdot x - 6 \cdot x \cdot x + 4 \cdot x - 0.1 \qquad (\mathbf{1})$$
$$P(x) = x(x(x - 6) + 4) - 0.1 \qquad (\mathbf{2})$$

Em (1), 5 multiplicações, 2 subtrações e 1 adição

Em (2), 3 multiplicações, 2 subtrações e 1 adição

Mesmo com menos operações, de (2) resulta:

$P(x) = x^3 - 6x^2 + 4x - 0.1$ pode ser escrito como:

$$P(x) = x \cdot x \cdot x - 6 \cdot x \cdot x + 4 \cdot x - 0.1 \qquad (\mathbf{1})$$
$$P(x) = x(x(x - 6) + 4) - 0.1 \qquad (\mathbf{2})$$

Em (1), 5 multiplicações, 2 subtrações e 1 adição

Em (2), 3 multiplicações, 2 subtrações e 1 adição

Mesmo com menos operações, de (2) resulta:

$$P(5.24) = 5.24(5.24(5.24 - 6) + 4) - 0.1$$
$P(5.24) = 0.005$ (sinal errado da aproximação !)

$P(5.24) = -0.00776$ (**valor exato**).

Outros exemplos de geração e propagação de erros

Outros exemplos de geração e propagação de erros

- Cálculos envolvendo somatórios:

$$S_k = \sum_{k=1}^{n} a_k,$$

onde os termos a_k podem ser positivos ou negativos.

Outros exemplos de geração e propagação de erros

- Cálculos envolvendo somatórios:

  $$S_k = \sum_{k=1}^{n} a_k,$$

  onde os termos a_k podem ser positivos ou negativos.

- Determinação numérica de integrais:

  $$y_n = \int_0^1 \frac{x^n}{x+a},$$

  para um valor fixo de a (a > 1) e, $n = 0, 1, 2, 3, \cdots$

Outros exemplos de geração e propagação de erros

- Cálculos envolvendo somatórios:

$$S_k = \sum_{k=1}^{n} a_k,$$

onde os termos a_k podem ser positivos ou negativos.

- Determinação numérica de integrais:

$$y_n = \int_0^1 \frac{x^n}{x+a},$$

para um valor fixo de a (a > 1) e, $n = 0, 1, 2, 3, \cdots$

A perda de dígitos significativos requer cuidados para evitar impacto negativo na computação em precisão finita.

- **Instabilidade Numérica**

Exemplo (algoritmo instável): Se $4AC << B^2$, a fórmula quadrática **não é conveniente** para a determinação da **menor raiz** da equação $Ax^2 + Bx + C = 0$.

Exemplo (algoritmo instável): Se $4AC << B^2$, a fórmula quadrática **não é conveniente** para a determinação da **menor raiz** da equação $Ax^2 + Bx + C = 0$.

Neste caso, quando $B > 0$, é desejável substituir a fórmula

$$x_1 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \quad \text{por} \quad x_1 = \frac{-2C}{B + \sqrt{B^2 - 4AC}}$$

Exemplo (algoritmo instável): Se $4AC << B^2$, a fórmula quadrática **não é conveniente** para a determinação da **menor raiz** da equação $Ax^2 + Bx + C = 0$.

Neste caso, quando $B > 0$, é desejável substituir a fórmula

$$x_1 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \quad \text{por} \quad x_1 = \frac{-2C}{B + \sqrt{B^2 - 4AC}}$$

Ou seja, $x_1 = \frac{C}{A}\frac{1}{x_2}$, ou $x_1\, x_2 = \frac{C}{A}$

Vamos resolver a equação: $x^2 - 1634x + 2 = 0.$

Vamos resolver a equação: $x^2 - 1634x + 2 = 0$.

Segue que,

$$x = \frac{1634 \pm \sqrt{1634^2 - 8}}{2} = 817 \pm \sqrt{667487}.$$

Vamos resolver a equação: $x^2 - 1634x + 2 = 0$.

Segue que,

$$x = \frac{1634 \pm \sqrt{1634^2 - 8}}{2} = 817 \pm \sqrt{667487}.$$

Temos então que:

$x_1 = 817 - 816.9987760 = 0.1224000000 \times 10^{-2}$

$x_2 = 817 + 816.9987760 = 0.1633998776 \times 10^{3}$

Note que os seis zeros da mantissa de x_1 resultam do cancelamento e portanto **não têm significado algum**.

Vamos resolver a equação: $x^2 - 1634x + 2 = 0$.

Segue que,
$$x = \frac{1634 \pm \sqrt{1634^2 - 8}}{2} = 817 \pm \sqrt{667487}.$$

Temos então que:
$x_1 = 817 - 816.9987760 = 0.1224000000 \times 10^{-2}$
$x_2 = 817 + 816.9987760 = 0.1633998776 \times 10^{3}$
Note que os seis zeros da mantissa de x_1 resultam do cancelamento e portanto **não têm significado algum**.

Mas usando, $x_1\, x_2 = \dfrac{C}{A}$ temos:

$x_1 = 0.1223991125 \times 10^{-2}$,
$x_1 = 0.1224000000 \times 10^{-2}$ (antes)

onde os dígitos da mantissa estão todos corretos.

Exemplo (fórmula de recorrência)

Se não utilizado da maneira adequada, os erros cometidos em uma **relação de recorrência** podem crescer exponencialmente e arruinar completamente os resultados.

Para calcular as integrais,

$$I_n = \int_0^1 \frac{x^n}{x+5}\,dx, \quad i = 1:N,$$

pode-se usar a relação de recorrência:

$$I_n + 5I_{n-1} = \frac{1}{n},$$

que segue diretamente do cálculo:

$$I_n + 5I_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5}\,dx = \int_0^1 x^{n-1}\,dx = \frac{1}{n}.$$

No que segue a fórmula acima será usada para calcular I_n, sempre com **seis** casas decimais. Para $n = 0$, temos:

$$I_0 = [\ln(x+5)]_0^1 \approx \ln 6 - \ln 5 = 0.182322.$$

Usando ainda a relação de recorrência para I_n, obtemos:

$$
\begin{aligned}
I_1 &= 1 - 5I_0 = 1 - 0.911610 = 0.088390, \\
I_2 &= 1/2 - 5I_1 = 0.500000 - 0.441950 = 0.058050, \\
I_3 &= 1/3 - 5I_2 = 0.333333 - 0.290250 = 0.043083, \\
I_4 &= 1/4 - 5I_3 = 0.250000 - 0.215415 = 0.034585, \\
I_5 &= 1/5 - 5I_4 = 0.200000 - 0.172925 = 0.027075, \\
I_6 &= 1/6 - 5I_5 = 0.166667 - 0.135375 = 0.031292, \\
I_7 &= 1/7 - 5I_6 = 0.142857 - 0.156460 = -0.013603.
\end{aligned}
$$

Mas em vista da relação de recorrência,

$$I_n + 5I_{n-1} = \frac{1}{n},$$

temos que tais resultados da tabela são inesperados, pois:

1^o) $I_6 > I_5$, e

2^o) $I_7 < 0$, que obviamente é absurdo !

A razão para o resultado absurdo acima é que o erro de arredondamento ϵ em $I_0 = 0.18232156...$, cuja magnitude é da ordem de 0.44×10^{-6}, é multiplicado por -5 no cálculo de I_1, que tem então um erro de -5ϵ

Este erro por sua vez produz um erro em I_2 de $5^2\epsilon$, e assim por diante. Segue que a magnitude do erro em I_7 é $5^7\epsilon = 0.0391$, que é maior do que o verdadeiro valor de I_7

Se uma maior precisão de máquina for utilizada, o resultado absurdo certamente vai aparecer em uma etapa posterior

Por exemplo, usando um computador que funciona com uma precisão que corresponde a cerca de 16 casas decimais forneceu um valor negativo para I_2, embora que I_0 tenha sido usado com toda precisão

Note que o efeito anterior pode ocorrer de forma similar para outras máquinas (verifique em seu computador!)

O algoritmo acima que faz uso de uma relação de recorrência é um exemplo de um fenômeno desagradável em cálculos computacionais, chamado de **instabilidade numérica**

Neste exemplo, pode-se evitar a instabilidade numérica, invertendo a direção da recursão. Usando a relação de recorrência na outra direção,

$$I_{n-1} = (1/n - I_n)/5,$$

observa-se que os erros serão divididos por um fator de 5 em cada passo. **Mas** note que ainda precisamos de um valor inicial para I_n !

Neste caso podemos ver diretamente a partir da definição

$$I_n + 5I_{n-1} = \frac{1}{n},$$

que I_n **diminui** com o aumento de n. Pode-se supor que I_n diminui lentamente quando n é grande (verifique).

Assim, podemos tentar $I_{12} = I_{11}$:

$$I_{11} + 5I_{11} \approx 1/12, \qquad I_{11} \approx 1/72 \approx 0.013889$$

o que mostra, $0 < I_{12} < 1/72 < I_{11}$. Usando ainda a relação de recorrência obtemos:

$$I_{10} = (1/11 - 0.013889)/5 = 0.015404, \qquad I_9 = (1/10 - 0.015404)/5 = 0.016919$$

$$I_8 = 0.018838, \quad I_7 = 0.021232, \quad I_6 = 0.024325, \quad I_5 = 0.028468,$$
$$I_4 = 0.034306, \quad I_3 = 0.043139, \quad I_2 = 0.058039, \quad I_1 = 0.088392,$$

e finalmente $I_0 = 0.182322$ (o valor correto!)

Exercício: Derive as relações de recorrência “avançada” e “recuada” para o cálculo das integrais,

$$I_n = \int_0^1 \frac{x^n}{4x+1}\,dx.$$

Diferentemente do exemplo anterior, explique a razão pela qual a recorrência é estável na direção “avançada” e instável na direção “recuada” ?

Um outro exemplo de fórmula de recorrência

Resolver a integral,

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

Solução: Vamos tentar encontrar uma fórmula de recorrência para I_n.

Integrando por partes (do cálculo), temos:

$$I_n = e^{-1} \left\{ [x^n e^x]_0^1 - \int_0^1 n x^{n-1} e^x \, dx \right\}$$

Exemplo (fórmula de recorrência)

Resolver a integral,

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

Solução: Vamos tentar encontrar uma fórmula de recorrência para I_n.

Integrando por partes (do cálculo), temos:

$$I_n = e^{-1} \left\{ [x^n e^x]_0^1 - \int_0^1 n x^{n-1} e^x \, dx \right\}$$

$$I_n = 1 - n \left(e^{-1} \int_0^1 x^{n-1} e^x \, dx \right)$$

Exemplo (fórmula de recorrência)

Resolver a integral,

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

Solução: Vamos tentar encontrar uma fórmula de recorrência para I_n.

Integrando por partes (do cálculo), temos:

$$I_n = e^{-1} \left\{ [x^n e^x]_0^1 - \int_0^1 n x^{n-1} e^x \, dx \right\}$$

$$I_n = 1 - n \left(e^{-1} \int_0^1 x^{n-1} e^x \, dx \right)$$

$$I_n = 1 - n I_{n-1}$$

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

E como sabemos calcular I_0, i.e.,

$$I_0 = e^{-1} \int_0^1 e^x \, dx = e^{-1}(e - 1) = 0.6321$$

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

E como sabemos calcular I_0, i.e.,

$$I_0 = e^{-1} \int_0^1 e^x \, dx = e^{-1}(e - 1) = 0.6321,$$

podemos, teoricamente, calcular I_n, usando $I_n = 1 - nI_{n-1}$.

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

E como sabemos calcular I_0, i.e.,

$$I_0 = e^{-1} \int_0^1 e^x \, dx = e^{-1}(e - 1) = 0.6321,$$

podemos, teoricamente, calcular I_n, usando $I_n = 1 - nI_{n-1}$.

Atenção !!

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

E como sabemos calcular I_0, i.e.,

$$I_0 = e^{-1} \int_0^1 e^x \, dx = e^{-1}(e - 1) = 0.6321,$$

podemos, teoricamente, calcular I_n, usando $I_n = 1 - nI_{n-1}$.

Atenção !!

Uma análise releva que a sequência I_n é **decrescente**!

A fórmula de recorrência é então dada por:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

E como sabemos calcular I_0, i.e.,

$$I_0 = e^{-1} \int_0^1 e^x \, dx = e^{-1}(e - 1) = 0.6321,$$

podemos, teoricamente, calcular I_n, usando $I_n = 1 - nI_{n-1}$.

Atenção !!

Uma análise releva que a sequência I_n é **decrescente**!

Realizando as contas de forma numérica $\cdots$

n	$I_n = 1 - nI_{n-1}$
0	$I_0 = 0.6321$
1	$I_1 = 0.3679$
2	$I_2 = 0.2642$
3	$I_3 = 0.2074$
4	$I_4 = 0.1704$
5	$I_5 = 0.1480$
6	$I_6 = 0.1120$
7	$I_7 = 0.2160$!!

Tabela 1: A sequência I_n é **decrescente**

De fato, note que:

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

De fato, note que:

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

$$I_n < e^{-1} \max_{0 \leq x \leq 1} \{e^x\} \int_0^1 x^n \, dx$$

De fato, note que:

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

$$I_n < e^{-1} \max_{0 \leq x \leq 1} \{e^x\} \int_0^1 x^n \, dx$$

$$I_n < e^{-1} \max_{0 \leq x \leq 1} \{e^x\} \int_0^1 x^n \, dx < \frac{1}{n+1}$$

De fato, note que:

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

$$I_n < e^{-1} \max_{0 \le x \le 1} \{e^x\} \int_0^1 x^n \, dx$$

$$I_n < e^{-1} \max_{0 \le x \le 1} \{e^x\} \int_0^1 x^n \, dx < \frac{1}{n+1}, \text{ Ou seja,}$$

$$I_7 < \frac{1}{8} < 0.1250$$

De fato, note que:

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx$$

$$I_n < e^{-1} \max_{0 \leq x \leq 1} \{e^x\} \int_0^1 x^n \, dx$$

$$I_n < e^{-1} \max_{0 \leq x \leq 1} \{e^x\} \int_0^1 x^n \, dx < \frac{1}{n+1}, \text{ Ou seja,}$$

$$I_7 < \frac{1}{8} < 0.1250$$

Vamos estudar um pouco mais a instabilidade numérica neste exemplo.

Pergunta: Como encontrar o valor exato para

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx \; ?$$

Pergunta: Como encontrar o valor exato para

$$I_n = e^{-1} \int_0^1 x^n e^x \, dx \; ?$$

É possível ? Tem alguma solução alternativa ?

Pergunta: Como encontrar o valor exato para $I_n = e^{-1} \int_0^1 x^n e^x \, dx$?

É possível ? Tem alguma solução alternativa ?

Para $I_n = e^{-1} \int_0^1 x^n e^x \, dx$, a resposta é sim!

Pergunta: Como encontrar o valor exato para $I_n = e^{-1} \int_0^1 x^n e^x \, dx$?

É possível ? Tem alguma solução alternativa ?

Para $I_n = e^{-1} \int_0^1 x^n e^x \, dx$, a resposta é sim!

Importante: uma relação de recorrência ser numericamente instável na **direção crescente** não impede de ser **estável na direção decrescente** de n.

Considere novamente a fórmula de recorrência

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

Considere novamente a fórmula de recorrência

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

Resolvendo para I_{n-1}, obtemos:

$$I_{n-1} = \frac{1 - I_n}{n}, \; n - 1, n - 2, n - 3, \cdots$$

Considere novamente a fórmula de recorrência

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

Resolvendo para I_{n-1}, obtemos:

$$I_{n-1} = \frac{1 - I_n}{n}, \; n - 1, n - 2, n - 3, \cdots$$

Note que a relação acima precisa de um valor inicial I_n.

Considere novamente a fórmula de recorrência

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

Resolvendo para I_{n-1}, obtemos:

$$I_{n-1} = \frac{1 - I_n}{n}, \; n - 1, n - 2, n - 3, \cdots$$

Note que a relação acima precisa de um valor inicial I_n.

Mas não é simples encontrar esse valor, pois todo I_n, onde $n > 0$, é desconhecido.

Considere novamente a fórmula de recorrência

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$

Resolvendo para I_{n-1}, obtemos:

$$I_{n-1} = \frac{1 - I_n}{n}, \; n - 1, n - 2, n - 3, \cdots$$

Note que a relação acima precisa de um valor inicial I_n.

Mas não é simples encontrar esse valor, pois todo I_n, onde $n > 0$, é desconhecido.

O que fazer ?

Do cálculo sabemos que $I_n \longrightarrow 0$ quando $n \longrightarrow \infty$

Do cálculo sabemos que $I_n \longrightarrow 0$ quando $n \longrightarrow \infty$

Assim, fazendo $I_{20} = 0$ e calculando I_{19}, I_{18}, I_{17}, $\cdots$

Do cálculo sabemos que $I_n \longrightarrow 0$ quando $n \longrightarrow \infty$

Assim, fazendo $I_{20} = 0$ e calculando I_{19}, I_{18}, I_{17}, $\cdots$

Obtêm-se $I_7 = 0.1123835 < 0.125$ (estimativa teórica !)

Do cálculo sabemos que $I_n \longrightarrow 0$ quando $n \longrightarrow \infty$

Assim, fazendo $I_{20} = 0$ e calculando I_{19}, I_{18}, I_{17}, $\cdots$

Obtêm-se $I_7 = 0.1123835 < 0.125$ (estimativa teórica)

Sendo $I_7 = 0$, obtêm-se $I_0 = 0.6320$ (0.6321 exato !)

Exercício. Vamos supor que:

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Atenção! Matematicamente temos:

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Atenção! Matematicamente temos:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$ (exato)

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Atenção! Matematicamente temos:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$ (exato)

$\tilde{I}_n = 1 - n\tilde{I}_{n-1}$, $n = 1, 2, 3, \cdots$ (calculado)

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Atenção! Matematicamente temos:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$ (exato)

$\tilde{I}_n = 1 - n\tilde{I}_{n-1}$, $n = 1, 2, 3, \cdots$ (calculado)

$\tilde{I}_0 = I_0 + \epsilon$ ($\epsilon =$ erro no valor inicial)

Exercício. Vamos supor que:

I_0 seja afetado por um erro.

Todas as operações aritméticas subsequentes são exatas.

Denotando $\cdots$

I_n como o valor exato, e

$\tilde{I}_n$ o valor calculado (com erro no valor inicial)

Atenção! Matematicamente temos:

$I_n = 1 - nI_{n-1}$, $n = 1, 2, 3, \cdots$ (exato)

$\tilde{I}_n = 1 - n\tilde{I}_{n-1}$, $n = 1, 2, 3, \cdots$ (calculado)

$\tilde{I}_0 = I_0 + \epsilon$ ($\epsilon =$ erro no valor inicial)

Continuando $\cdots$

Defina o erro r_n como:

$r_n = \tilde{I}_n - I_n$, $n = 1, 2, 3, \cdots$

Defina o erro r_n como:

$r_n = \tilde{I}_n - I_n$

$r_n = -nr_{n-1}$, $n = 1, 2, 3, \cdots$

Defina o erro r_n como:

$r_n = \tilde{I}_n - I_n$

$r_n = -nr_{n-1}$, $n = 1, 2, 3, \cdots$

Agora aplique repetidamente a fórmula acima para obter uma relação de recorrência para r_n em termos do erro ϵ, que é introduzido no cálculo do valor inicial I_0.

(i) Com o resultado acima, explique a fonte do acúmulo do erro da relação de recorrência na direção “avançada”.

(ii) Ainda com base na análise acima, explique também a estabilidade da relação de recorrência na outra direção (“recuada”).

- **Mal condicionamento**

Conceitos de Cálculo Numérico (revisitando)

O objetivo deste curso é examinar alguns métodos para a resolução **numérica** de vários tipos de problemas

Queremos implementar algoritmos em um computador que levará à solução (ou uma aproximação razoável) do problema

Qualquer problema pode ser pensado como uma função dos dados de entrada:

x que entra

uma função f é executada

$f(x)$ que sai

O condicionamento é um conceito que se aplica aos problemas

Pergunta: Quais são os efeitos que pequenas alterações nos dados de entrada têm sobre a solução do problema ?

Um problema está **bem condicionado** se uma pequena variação (na entrada) produz uma pequena alteração em $f(x)$ (na saída)

Um problema é **mal condicionado** se uma pequena alteração na entrada produz uma grande mudança na saída (dizemos que esses problemas são sensíveis a pequenas perturbações nos dados)

Por exemplo, o número de condição pode estar associado a sistemas lineares de equações

Além disso, os números de condição podem estar associados a outros problemas, tais como:

Cálculo de raízes de polinômios

Cálculo de autovalores e autovetores

Resolução de equações diferenciais ordinárias e parciais

O que queremos dizer sobre um sistema linear de equações ser **mal condicionado** ou **bem condicionado** ?

O que queremos dizer sobre um sistema linear de equações ser **mal condicionado** ou **bem condicionado** ?

Um sistema de equações é considerado **bem condicionado** se uma **pequena** mudança na **matriz de coeficientes** ou uma **pequena** mudança no lado direito **resulta** em uma **pequena** mudança no vetor solução

O que queremos dizer sobre um sistema linear de equações ser **mal condicionado** ou **bem condicionado** ?

Um sistema de equações é considerado **bem condicionado** se uma **pequena** mudança na **matriz de coeficientes** ou uma **pequena** mudança no lado direito **resulta** em uma **pequena** mudança no vetor solução

Um sistema de equações é considerado **mal condicionado** se uma **pequena** mudança na **matriz de coeficientes** ou uma **pequena** mudança no lado direito **resulta** em uma **grande** mudança no vetor solução

EXEMPLO 1

EXEMPLO 1

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

EXEMPLO 1

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

SOLUÇÃO

EXEMPLO 1

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

SOLUÇÃO

A solução para o conjunto de equações acima é:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

EXEMPLO 1

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

SOLUÇÃO

A solução para o conjunto de equações acima é:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4 \end{bmatrix}$$

Faça uma pequena mudança na matriz dos coeficientes das equações

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.998 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4 \end{bmatrix}$$

Faça uma pequena mudança na matriz dos coeficientes das equações

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.998 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

A solução agora muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6.989 \\ -1.497 \end{bmatrix}$$

Assim, o sistema de equações original (abaixo) “parece” ser **mal condicionado**, porque uma pequena mudança na **matriz de coeficientes** ou do **lado direito** RESULTOU em uma **grande** mudança no vetor solução

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

EXEMPLO 2

EXEMPLO 2

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

EXEMPLO 2

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

SOLUÇÃO

EXEMPLO 2

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

SOLUÇÃO

A solução para o conjunto de equações acima é:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

EXEMPLO 2

Este sistema de equações é bem condicionado ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

SOLUÇÃO

A solução para o conjunto de equações acima é:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}$$

Faça uma pequena mudança na matriz dos coeficientes das equações

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

A solução muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}$$

Faça uma pequena mudança na matriz dos coeficientes das equações

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

A solução agora muda para:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.003 \\ 0.997 \end{bmatrix}$$

Assim, o sistema de equações original (abaixo) “parece” ser **bem condicionado**, porque uma pequena mudança na **matriz de coeficientes** ou do **lado direito** NÃO RESULTOU em uma **grande** mudança no vetor solução

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

Assim, o sistema de equações original (abaixo) “parece” ser **bem condicionado**, porque uma pequena mudança na **matriz de coeficientes** ou do **lado direito** NÃO RESULTOU em uma **grande** mudança no vetor solução

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

Então, o que dizer se o sistema linear de equações for **mal condicionado** ou **bem condicionado** ?

Assim, o sistema de equações original (abaixo) “parece” ser **bem condicionado**, porque uma pequena mudança na **matriz de coeficientes** ou do **lado direito** NÃO RESULTOU em uma **grande** mudança no vetor solução

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

Então, o que dizer se o sistema linear de equações for **mal condicionado** ou **bem condicionado** ?

Bem, se um sistema linear de equações é mal condicionado, então não podemos confiar tanto na sua solução, quando resolvido por um procedimento numérico **QUALQUER** !

Pergunta: Mas podemos pelo menos quantificar quantos algarismos significativos se pode confiar na solução ?

Pergunta: Mas podemos pelo menos quantificar quantos algarismos significativos se pode confiar na solução ?

Resposta: Em princípio, sim. Mas, como ?

Pergunta: Mas podemos pelo menos quantificar quantos algarismos significativos se pode confiar na solução ?

Resposta: Em princípio, sim. Mas, como ?

Toda matriz inversível tem um número de condição, e juntamente com a epsilon da máquina, podemos quantificar quantos algarismos significativos se pode confiar na solução.

Pergunta: Mas podemos pelo menos quantificar quantos algarismos significativos se pode confiar na solução ?

Resposta: Em princípio, sim. Mas, como ?

Toda matriz inversível tem um número de condição e juntamente com a epsilon da máquina, podemos quantificar quantos algarismos significativos se pode confiar na solução

Para calcular o número de condição de uma matriz **quadrada inversível**, é preciso saber o que é uma norma para matrizes

Ou seja, como a norma de uma matriz é definida ?

Ou seja, como a norma de uma matriz é definida ?

Atenção! Lembrando $\cdots$

Ou seja, como a norma de uma matriz é definida ?

Atenção! Lembrando $\cdots$

Assim como o fator determinante, a norma de uma matriz é um número, um simples escalar

Ou seja, como a norma de uma matriz é definida ?

Atenção! Lembrando $\cdots$

Assim como o fator determinante, a norma de uma matriz é um número, um simples escalar

Exceto pela matriz nula, a norma de uma matriz é sempre positiva e é definida para todas as matrizes quadradas ou retangulares, e matrizes quadradas inversíveis ou não inversíveis

Uma norma para matriz tipicamente empregada é a **norma linha**:

$$||A||_\infty = \max_{1 \leq i \leq m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\},$$

onde A é uma matriz $m \times n$.

Uma norma para matriz tipicamente empregada é a **norma linha**:

$$||A||_\infty = \max_{1 \le i \le m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\},$$

onde A é uma matriz $m \times n$.

EXEMPLO 3

Encontre a norma linha da seguinte matriz A.

$$A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

$$||A||_\infty = \max\{10+7+0, 3+2.009+6, 5+1+5\}$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

$$||A||_\infty = \max\{10+7+0, 3+2.009+6, 5+1+5\}$$

$$||A||_\infty = \max\{17, 11.009, 11\}$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

$$||A||_\infty = \max\{10+7+0, 3+2.009+6, 5+1+5\}$$

$$||A||_\infty = \max\{17, 11.009, 11\}$$

$$||A||_\infty = 17$$

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

$$||A||_\infty = \max\{10+7+0, 3+2.009+6, 5+1+5\}$$

$$||A||_\infty = \max\{17, 11.009, 11\}$$

$$||A||_\infty = 17$$

Como a norma está relacionada com o condicionamento da matriz ?

SOLUÇÃO

$$||A||_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}, \text{ sendo } A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.009 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

$$||A||_\infty =$$

$$\max\{|10|+|-7|+|0|, |-3|+|2.009|+|6|, |5|+|-1|+|5|\}$$

$$||A||_\infty = \max\{10+7+0, 3+2.009+6, 5+1+5\}$$

$$||A||_\infty = \max\{17, 11.009, 11\}$$

$$||A||_\infty = 17$$

Como a norma está relacionada com o condicionamento da matriz ?

Vamos responder esta pergunta usando um exemplo !

Voltando ao sistema linear de equações mal condicionado,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

onde a solução é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Voltando ao sistema linear de equações mal condicionado,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

onde a solução é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Denotando o sistema acima por

$AX = C$, temos:

$||X||_\infty = 2$

$||C||_\infty = 7.999$

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

onde a solução agora é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

onde a solução agora é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Denotando o sistema acima por

$AX' = C'$,

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

onde a solução agora é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Denotando o sistema acima por

$AX' = C'$,

a mudança no vetor do lado direito é dado por

$\Delta C = C' - C$

Faça uma pequena mudança no vetor do lado direito das equações

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

onde a solução agora é dada por:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Denotando o sistema acima por
$AX' = C'$,
a mudança no vetor do lado direito é dado por
$\Delta C = C' - C$

e a mudança no vetor solução é dado por
$\Delta X = X' - X$

Segue que,

$$\Delta C = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix} - \begin{bmatrix} 4 \\ 7.999 \end{bmatrix} = \begin{bmatrix} 0.001 \\ -0.001 \end{bmatrix}$$

e

$$\Delta X = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -5.999 \\ 3.000 \end{bmatrix}$$

Segue que,

$$\Delta C = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix} - \begin{bmatrix} 4 \\ 7.999 \end{bmatrix} = \begin{bmatrix} 0.001 \\ -0.001 \end{bmatrix}$$

e

$$\Delta X = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -5.999 \\ 3.000 \end{bmatrix}$$

então,

$$||\Delta C||_\infty = 0.001$$

$$||\Delta X||_\infty = 5.999$$

Segue que,

$$\Delta C = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix} - \begin{bmatrix} 4 \\ 7.999 \end{bmatrix} = \begin{bmatrix} 0.001 \\ -0.001 \end{bmatrix}$$

e

$$\Delta X = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -5.999 \\ 3.000 \end{bmatrix}$$

então,

$$||\Delta C||_\infty = 0.001$$

$$||\Delta X||_\infty = 5.999$$

Atenção !

A mudança relativa na norma do vetor solução é

$$\frac{||\Delta X||_\infty}{||X||_\infty} = \frac{5.999}{2} = 2.9995$$

A mudança relativa na norma no vetor do lado direito é

$$\frac{||\Delta C||_\infty}{||C||_\infty} = \frac{0.001}{7.999} = 1.250 \times 10^{-4}$$

Notem que uma **pequena** mudança relativa de 1.250×10^{-4} no vetor do lado direito resulta em uma **grande** mudança (**ordem de grandeza**) no vetor solução de 2.9995

De fato, a razão entre a mudança relativa na norma do vetor solução e a mudança relativa na norma do vetor do lado direito é

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{2.9995}{1.250 \times 10^{-4}} = 23996$$

De fato, a razão entre a mudança relativa na norma do vetor solução e a mudança relativa na norma do vetor do lado direito é

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{2.9995}{1.250 \times 10^{-4}} = 23996$$

Repetindo as mesmas contas para o sistema bem condicionando $\cdots$

De fato, a razão entre a mudança relativa na norma do vetor solução e a mudança relativa na norma do vetor do lado direito é

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{2.9995}{1.250 \times 10^{-4}} = 23996$$

Repetindo as mesmas contas para o sistema bem condicionando $\cdots$

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{5 \times 10^{-4}}{1.429 \times 10^{-4}} \approx 3.5$$

De fato, a razão entre a mudança relativa na norma do vetor solução e a mudança relativa na norma do vetor do lado direito é

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{2.9995}{1.250 \times 10^{-4}} = 23996$$

Repetindo as mesmas contas para o sistema bem condicionando $\cdots$

$$\frac{||\Delta X||_\infty / ||X||_\infty}{||\Delta C||_\infty / ||C||_\infty} = \frac{5 \times 10^{-4}}{1.429 \times 10^{-4}} \approx 3.5$$

Algumas perguntas $\cdots$

Existe alguma relação geral entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e/ou} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \ ?$$

Existe alguma relação geral entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e/ou} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \ ?$$

Se existe, tal relação pode ajudar a identificar sistemas lineares de equações que são **mal condicionados** e **bem condicionados** ?

Existe alguma relação geral entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e/ou} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \ ?$$

Se existe, tal relação pode ajudar a identificar sistemas lineares de equações que são **mal condicionados** e **bem condicionados** ?

Ou seja, se tal relação existe, será possível quantificar o condicionamento de matrizes associadas a sistemas lineares de equações ?

Existe alguma relação geral entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e/ou} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \ ?$$

Se existe, tal relação pode ajudar a identificar sistemas lineares de equações que são **mal condicionados** e **bem condicionados** ?

Ou seja, se tal relação existe, será possível quantificar o condicionamento de matrizes associadas a sistemas lineares de equações ?

Mais do que isso, se tal relação existe, será possível quantificar quantos dígitos significativos podemos confiar na solução numérica de um sistema linear de equações ?

Resposta: Sim, existe relações gerais entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \text{ !!}$$

Resposta: Sim, existe relações gerais entre

$$\frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta C||}{||C||} \quad \textbf{e} \quad \frac{||\Delta X||}{||X||} \text{ e } \frac{||\Delta A||}{||A||} \text{ !!}$$

A relação entre $\frac{||\Delta X||}{||X||}$ e $\frac{||\Delta C||}{||C||}$, é dada por:

$$\frac{||\Delta X||}{||X||} \leq ||A||\,||A^{-1}||\frac{||\Delta C||}{||C||}$$

Resposta: Sim, existe relações gerais entre $\frac{||\Delta X||}{||X||}$ e $\frac{||\Delta C||}{||C||}$ **e** $\frac{||\Delta X||}{||X||}$ e $\frac{||\Delta A||}{||A||}$!!

A relação entre $\frac{||\Delta X||}{||X||}$ e $\frac{||\Delta C||}{||C||}$, é dada por:

$$\frac{||\Delta X||}{||X||} \leq ||A||\,||A^{-1}||\frac{||\Delta C||}{||C||}$$

E a relação entre $\frac{||\Delta X||}{||X||}$ e $\frac{||\Delta A||}{||A||}$, é dada por:

$$\frac{||\Delta X||}{||X+\Delta X||} \leq ||A||\,||A^{-1}||\frac{||\Delta A||}{||A||}$$

As duas desigualdades revelam que a mudança relativa na norma do vetor do lado direito **ou** na matriz de coeficientes pode ser amplificada pelo **produto** por $||A||\,||A^{-1}||$

As duas desigualdades revelam que a mudança relativa na norma do vetor do lado direito **ou** na matriz de coeficientes pode ser amplificada pelo **produto** por $||A||\,||A^{-1}||$

O **número** $||A||\,||A^{-1}||$ é chamado número de condição da matriz; i.e., $Cond(A) \equiv ||A||\,||A^{-1}||$

As duas desigualdades revelam que a mudança relativa na norma do vetor do lado direito **ou** na matriz de coeficientes pode ser amplificada pelo **produto** por $||A||\,||A^{-1}||$

O **número** $||A||\,||A^{-1}||$ é chamado número de condição da matriz; i.e., $Cond(A) \equiv ||A||\,||A^{-1}||$

O desejável é que $Cond(A) \equiv ||A||\,||A^{-1}|| \approx 1$

As duas desigualdades revelam que a mudança relativa na norma do vetor do lado direito **ou** na matriz de coeficientes pode ser amplificada pelo **produto** por $||A||\,||A^{-1}||$

O **número** $||A||\,||A^{-1}||$ é chamado número de condição da matriz; i.e., $||A||\,||A^{-1}|| \equiv Cond(A)$

O desejável é que $Cond(A) \equiv ||A||\,||A^{-1}|| \approx 1$

Além disso, o número de condição da matriz $Cond(A)$, em conjunto com o valor epsilon de máquina "ϵ_{maq}", pode ser usado para quantificar (em teoria) a precisão do número de dígitos significativos para a solução numérica de $AX = C$

Então, como podemos utilizar os resultados anteriores para estimar quantos dígitos significativos estão corretos para a solução numérica de $AX = C$?

Então, como podemos utilizar os resultados anteriores para estimar quantos dígitos significativos estão corretos para a solução numérica de $AX = C$?

Lembre que

$$\frac{||\Delta X||}{||X||} = \frac{||X' - X||}{||X||} \leq Cond(A) \times \frac{||\Delta C||}{||C||}$$

Então, como podemos utilizar os resultados anteriores para estimar quantos dígitos significativos estão corretos para a solução numérica de $AX = C$?

Lembre que

$$\frac{||\Delta X||}{||X||} = \frac{||X' - X||}{||X||} \leq Cond(A) \times \frac{||\Delta C||}{||C||}$$

Daí, o possível erro relativo no vetor solução é dado por:

$$\text{“}possível\ erro\ relativo\ no\ vetor\ X\text{”} \leq Cond(A) \times \epsilon_{maq}$$

Assim, $Cond(A) \times \epsilon_{maq}$ deve fornecer o número de dígitos significativos, pelo menos m dígitos de confiança na solução, se comparado com $\frac{1}{2} \times 10^m$

Assim, $Cond(A) \times \epsilon_{maq}$ deve fornecer o número de dígitos significativos, pelo menos m dígitos de confiança na solução, se comparado com $\frac{1}{2} \times 10^m$

Vejamos alguns exemplos para ajudar a fixar as ideias

Assim, $Cond(A) \times \epsilon_{maq}$ deve fornecer o número de dígitos significativos, pelo menos m dígitos de confiança na solução, se comparado com $\frac{1}{2} \times 10^m$

Vejamos alguns exemplos para ajudar a fixar as ideias

EXEMPLO

Quantos dígitos significativos podemos confiar na solução do seguinte sistema de equações ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

Assim, $Cond(A) \times \epsilon_{maq}$ deve fornecer o número de dígitos significativos, pelo menos m dígitos de confiança na solução, se comparado com $\frac{1}{2} \times 10^m$

Vejamos alguns exemplos para ajudar a fixar as ideias

EXEMPLO 1

Quantos dígitos significativos podemos confiar na solução do seguinte sistema de equações ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

SOLUÇÃO

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$$||A||_\infty = 5.999 \text{ e } ||A^{-1}||_\infty = 5999$$

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$$||A||_\infty = 5.999 \text{ e } ||A^{-1}||_\infty = 5999$$

$$Cond(A) = ||A||_\infty \, ||A^{-1}||_\infty \approx 35988$$

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$||A||_\infty = 5.999$ e $||A^{-1}||_\infty = 5999$

$Cond(A) = ||A||_\infty \, ||A^{-1}||_\infty \approx 35988$

Supondo uma precisão simples com 24 bits na mantissa, o "epsilon de máquina" ($\epsilon_{maq} = 2^{1-24} = 0.119209 \times 10^{-6}$)

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$$||A||_\infty = 5.999 \text{ e } ||A^{-1}||_\infty = 5999$$

$$Cond(A) = ||A||_\infty \, ||A^{-1}||_\infty \approx 35988$$

Supondo uma precisão simples com 24 bits na mantissa, o "epsilon de máquina" ($\epsilon_{maq} = 2^{1-24} = 0.119209 \times 10^{-6}$)

$$Cond(A) \times \epsilon_{maq} = 35988 \times 0.119209 \times 10^{-6} = 0.429 \times 10^{-2}$$

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$||A||_\infty = 5.999$ e $||A^{-1}||_\infty = 5999$

$Cond(A) = ||A||_\infty \, ||A^{-1}||_\infty \approx 35988$

Supondo uma precisão simples com 24 bits na mantissa, o "epsilon de máquina" ($\epsilon_{maq} = 2^{1-24} = 0.119209 \times 10^{-6}$)

$$Cond(A) \times \epsilon_{maq} = 35988 \times 0.119209 \times 10^{-6} = 0.429 \times 10^{-2}$$

Daí, $\frac{1}{2} \times 10^m \leq 0.429 \times 10^{-2}$ (dois dígitos de confiança)

EXEMPLO 2

Quantos dígitos significativos podemos confiar na solução do seguinte sistema de equações ?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

SOLUÇÃO

Para este sistema, temos que

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} -3 & 2 \\ 2 & -1 \end{bmatrix}$$

$||A||_\infty = 5$ e $||A^{-1}||_\infty = 5$

$Cond(A) = ||A||_\infty \, ||A^{-1}||_\infty = 25$

Supondo uma precisão simples com 24 bits na mantissa, o "epsilon de máquina" ($\epsilon_{maq} = 2^{1-24} = 0.119209 \times 10^{-6}$)

$Cond(A) \times \epsilon_{maq} = 25 \times \epsilon_{maq} = 2.980225 \times 10^{-6}$

Daí, $\frac{1}{2} \times 10^m \leq 0.298023 \times 10^{-5}$ (cinco dígitos de confiança !)

Ainda sobre bem ou mal condicionamento $\cdots$

Suponha agora que o problema consiste simplesmente em avaliar a função f (de uma variável real) em um ponto x.

Pergunta: Se x é ligeiramente perturbado, qual é o efeito sobre $f(x)$?

Se a pergunta refere-se ao **erro absoluto**, pode-se então invocar o Teorema do valor médio e escrever:

$$f'(\xi) = \frac{f(x+h) - f(x)}{h}, \text{ ou melhor}$$

$$f(x+h) - f(x) = f'(\xi)h, \quad x < \xi < x+h$$

Ainda sobre bem ou mal condicionamento $\cdots$

Desta maneira, se $f'(x)$ não é muito grande, o efeito da perturbação sobre $f(x)$ é pequeno

Usualmente, entretanto, é o **erro relativo** que fornece uma medida com maior significado para tais questões.

Em pertubar x por uma quantidade h, tem-se que a quantidade $\dfrac{h}{x}$ como o relativo tamanho da pertubação.

Da mesma forma, quando $f(x)$ é pertubado para $f(x+h)$, o tamanho relativo dessa pertubação é:

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{h\,f'(x)}{f(x)} = \left[\frac{x\,f'(x)}{f(x)}\right] \left(\frac{h}{x}\right)$$

Desta maneira, o fator $\left[\dfrac{x\,f'(x)}{f(x)}\right]$ serve com o número de condição para este problema

EXEMPLO: Qual é o número de condição para o cálculo da função inversa do seno ?

SOLUÇÃO:

Seja $f(x) = sen^{-1} = arcsen\, x$

Segue que, $\dfrac{x\,f'(x)}{f(x)} = \dfrac{x}{\sqrt{1-x^2}\,sen^{-1}}$

Para x próximo de 1, $sen^{-1} \approx \pi/2$.

Desta maneira, o fator $\left[\frac{x\,f'(x)}{f(x)}\right]$

O número de condição se torna infinito na medida em que x se aproxima de 1, uma vez que o número de condição é aproximado por:

$$\frac{2x}{\pi\sqrt{1-x^2}}$$

Assim, pequenos erros relativos em x podem conduzir para grandes erros relativos em $f(x) = sen^{-1}$, para $x \approx 1$

- **“Desastres numéricos” (ou falha humana ?)**

  Ver alguns exemplos em

  http://www.ime.unicamp.br/∼ms211/material-didático

Sumário

Nenhum método numérico pode compensar um problema mal condicionado

Mas nem todo método numérico será bom para um problema bem condicionado

Um método numérico precisa controlar os diversos erros computacionais (e.g., aproximação, o truncamento, o arredondamento, que se propagam), equilibrando o custo (tempo) computacional final

Um método numérico deve ser consistente e estável de modo a convergir para a resposta correta

O padrão IEEE (IEEE - Institute of Electrical and Electronic Engineers) “tenta” padronizar a precisão simples e dupla em ponto flutuante e a sua aritmética
IEEE websites:

http://www.ieee.org.br (**Brasil**)
http://www.ieee.org (**Internacional**)

Visite também o website:

http://grouper.ieee.org/groups/754/ (**Internacional**)

Overflow e *underflow* numéricos, e cancelamento, devem ser cuidadosamente considerados e evitados ou contornados

Formas matematicamente equivalentes não são numericamente equivalentes!

Proposição: Seja A tal que $I = A\,A^{-1}$. Se $AX = C$, então

$$\frac{||\Delta X||}{||X + \Delta X||} \leq ||A||\,||A^{-1}||\frac{||\Delta A||}{||A||}$$

Prova: Seja $AX = C$. Se A é modificado para A', X será modificado para X', tal que:

$A'\,X' = C$

Das duas equações acima, temos que $AX = A'\,X'$

Denotando a mudança nas matrizes A e X por ΔA e ΔX, respectivamente,

$\Delta A = A' - A$ e $\Delta X = X' - X$.

Então, $AX = (A + \Delta A)(X + \Delta X)$

Expandindo a última equação:

$$\begin{aligned} AX &= (A + \Delta A)(X + \Delta X) \\ AX &= AX + A\Delta X + \Delta AX + \Delta A\Delta X \\ 0 &= A\Delta X + \Delta AX + \Delta A\Delta X \\ -A\Delta X &= \Delta A(X + \Delta X) \\ \Delta X &= -A^{-1}\Delta A(X + \Delta X) \end{aligned}$$

Aplicando o teorema "usual" das normas, onde estabelece que a norma do produto de duas matrizes é menor do que o produto das normas das matrizes, segue que:

$$||\Delta X|| \leq || - A^{-1}||\,||\Delta A||\,||X + \Delta X||$$

Multiplicando ambos os lados por $||A||\,||X + \Delta X||^{-1}$, fica:

$$\frac{||\Delta X||}{||X + \Delta X||} \leq ||A||\,||A^{-1}||\frac{||\Delta A||}{||A||}, \text{ ou ainda:}$$

$$\frac{||\Delta X||}{||X + \Delta X||} \leq ||A||\,||A^{-1}||\frac{||\Delta A||}{||A||}$$

Lembrete: Propriedade de normas de matrizes

- Para uma matriz A, $||A|| \geq 0$
- Para uma matriz A e um escalar k, $||k\,A|| = k||A||$
- Para duas matrizes A e B de mesma ordem, $||A + B|| \leq ||A|| + ||B||$
- Para duas matrizes A e B, que podem ser multiplicadas como $A\,B$, $\quad ||A\,B|| \leq ||A||\,||B||$