Capítulo 1 – Representação de Números e Erros

1.1. Introdução

Como aparecem os erros?

Quais os seus efeitos?

Como controlar esses efeitos?

Tipos de Erros:

- Erros inerentes à matematização do fenómeno físico: os sistemas adoptados para representar a realidade (modelos) são geralmente (necessariamente) aproximações;

- Erros nos dados: resultam da incerteza existente nas medições de grandezas físicas. Devem-se às precisões finitas e limitadas dos instrumentos de medida. São estudados no âmbito das Probabilidades e Estatística;

- Erros de método ou de truncatura: resultam de substituir o modelo matemático adoptado por um processo de tratamento numérico aproximado. Exemplos: substituir derivadas por razões incrementais, integrais por somatórios, séries por somas de um número finito de termos;

- Erros de arredondamento: surgem pelas limitações dos instrumentos de cálculo utilizados na efectivação de operações numéricas elementares, os quais trabalham com um número limitado de algarismos.

1.2. Valores aproximados e erros

1.2.1. Erro Absoluto e Erro Relativo

Definição: Erro Absoluto

Seja $\bar{p}$ um valor aproximado de p.

Chama-se **erro absoluto** do valor aproximado $\bar{p}$ a

$$\Delta_{\bar{p}} = |p - \bar{p}|$$

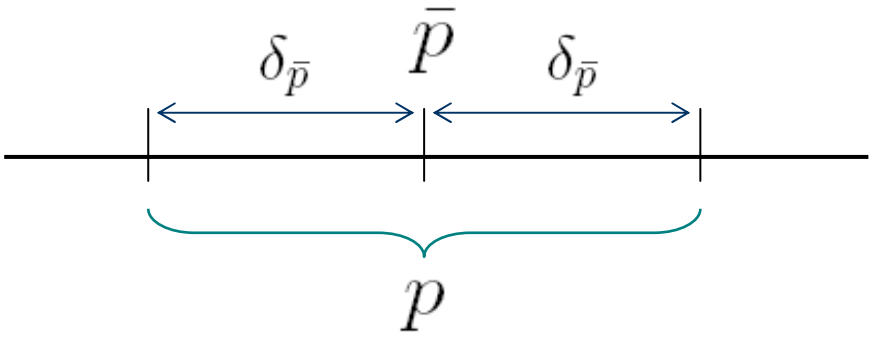
como este valor é geralmente desconhecido, faz mais sentido falar em:

Definição: Majorante do Erro Absoluto

Um **majorante do erro absoluto** do valor aproximado $\bar{p}$ é um valor $\delta_{\bar{p}}$ tal que

$$\delta_{\bar{p}} \geq \Delta_{\bar{p}}$$

e assim, $\bar{p} - \delta_{\bar{p}} \leq p \leq \bar{p} + \delta_{\bar{p}}$



Para uma melhor percepção da qualidade da aproximação, o valor do erro deve ser independente da ordem de grandeza, por isso:

Definição: Erro Relativo

Seja $\bar{p}$ um valor aproximado de $p \neq 0$.

Chama-se **erro relativo do valor aproximado** $\bar{p}$ a

$$r_{\bar{p}} = \frac{\Delta_{\bar{p}}}{|p|}$$

O Erro Relativo é portanto uma grandeza sem dimensões.

Ao produto $100\ r_{\bar{p}}$ chamamos **percentagem de erro**, expresso em percentagem.

Definição: Majorante do Erro Relativo

Um **majorante do erro relativo** do valor aproximado $\bar{p}$ é um valor $\varepsilon_{\bar{p}}$ tal que

$$\varepsilon_{\bar{p}} \geq r_{\bar{p}}$$

Alguns casos concretos:

Exemplo 1.4 *Seja* $x = 3.141592$ *e* $\overline{x} = 3.14$. *Neste caso,*

$$\Delta_{\overline{x}} = |x - \overline{x}| = |3.141592 - 3.14| = 0.001592$$

e

$$r_{\overline{x}} = \frac{\Delta_{\overline{x}}}{|x|} = \frac{0.001592}{3.141592} = 0.000507.$$

Neste caso, não existe muita diferença entre o erro absoluto e o erro relativo.

Considere-se agora $y = 1000000$ *e* $\overline{y} = 999996$. *Assim,*

$$\Delta_{\overline{y}} = |y - \overline{y}| = |1000000 - 999996| = 4$$

e

$$r_{\overline{y}} = \frac{\Delta_{\overline{y}}}{|y|} = \frac{4}{1000000} = 0.000004.$$

Como os valores são de grande magnitude, apesar do erro absoluto ser elevado, a aproximação pode ser considerada boa.

Finalmente, seja $z = 0.0000012$ *e* $\overline{z} = 0.000009$. *Agora*

$$\Delta_{\overline{z}} = |z - \overline{z}| = |0.0000012 - 0.000009| = 0.000003$$

e

$$r_{\overline{z}} = \frac{\Delta_{\overline{z}}}{|z|} = \frac{0.000003}{0.0000012} = 0.25.$$

Caso oposto ao anterior: um erro relativo de 25% não é aceitável.

Relações entre majorantes:

Caso 1: Sendo conhecido um **majorante do erro absoluto**, encontrar um **majorante para o erro relativo**.

$$\delta_{\bar{p}} \hookrightarrow \varepsilon_{\bar{p}}$$

A partir da definição do Erro Relativo, procuremos um majorante:

$$r_{\bar{p}} = \frac{\Delta_{\bar{p}}}{|p|}$$

majorando o numerador: $\delta_{\bar{p}} \geq \Delta_{\bar{p}}$

minorando o denominador: $|\overline{p}| - \delta_{\overline{p}} \leq |p|$

assim, uma estimativa de $\varepsilon_{\bar{p}} \geq r_{\bar{p}}$ é dada por:

$$\varepsilon_{\overline{p}} = \frac{\delta_{\overline{p}}}{|\overline{p}| - \delta_{\overline{p}}}$$

Em muitos casos, $\delta_{\overline{p}} \ll |\overline{p}|$ e então podemos simplificar:

$$\varepsilon_{\overline{p}} \approx \frac{\delta_{\overline{p}}}{|\overline{p}|}$$

Caso 2: Sendo conhecido um **majorante do erro relativo**, encontrar um **majorante para o erro absoluto**.

$$\varepsilon_{\overline{p}} \hookrightarrow \delta_{\overline{p}}$$

Consideremos a definição do Erro Relativo, escrita na forma:

$$\Delta_{\overline{p}} = r_{\overline{p}} \, |p|$$

Tratando-se de um produto, procuremos majorantes para ambos os factores:

$$\varepsilon_{\overline{p}} \geq r_{\overline{p}}$$

$$|p| \leq |\overline{p}| + \delta_{\overline{p}}$$

portanto, uma estimativa de $\delta_{\overline{p}} \geq \Delta_{\overline{p}}$ é dada por:

$$\delta_{\overline{p}} = \varepsilon_{\overline{p}} \left(|\overline{p}| + \delta_{\overline{p}} \right)$$

donde, assumindo que $\varepsilon_{\overline{p}} < 1$, resulta:

$$\delta_{\overline{p}} = \frac{\varepsilon_{\overline{p}}}{1 - \varepsilon_{\overline{p}}} \, |\overline{p}|$$

Dado um número aproximado e o seu erro, como identificar os algarismos significativos?

1.2.2. Algarismos Significativos

Definição: Algarismos Significativos

Seja $\bar{p}$ um valor aproximado de p. Diz-se que um algarismo de $\bar{p}$ é significativo se $\Delta_{\bar{p}} \leq 0.5 \times 10^{-m}$ onde m é a ordem decimal da casa que esse algarismo ocupa, nas seguintes condições:

- ser diferente de zero;
- ser igual a zero mas existir à sua esquerda um algarismo significativo diferente de zero.

Exemplo1: Seja $\bar{p} = 396.74$ com $\Delta_{\bar{p}} = 0.05$.

- o algarismo 4 não é significativo.
- o algarismo 7 é significativo.
- os restantes algarismos são significativos.

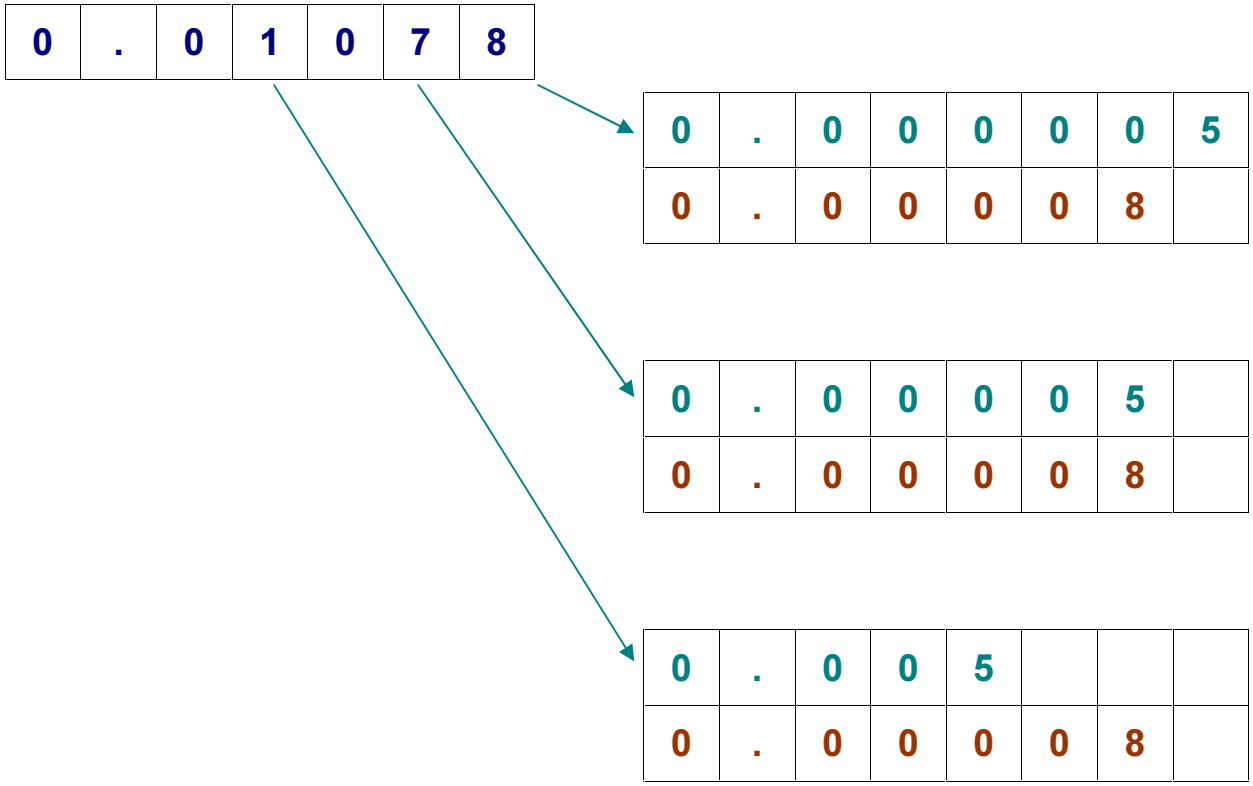
3	9	6	.	7	4

0	.	0	0	5
0	.	0	5	

0	.	0	5	
0	.	0	5	

Exemplo2: Seja $\bar{p} = 0.01078$ com $\Delta_{\bar{p}} = 0.00008$.

- o algarismo 8 não é significativo $(0.00008 > 0.5 \times 10^{-5})$.
- o algarismo 7 não é significativo $(0.00008 > 0.5 \times 10^{-4})$.
- o algarismo 0 (entre os algarismos 1 e 7) é significativo.
- o algarismo 1 é significativo $(0.00008 \leq 0.5 \times 10^{-2})$.
- os dois zeros à esquerda não são significativos.



1.3. Representação de números

Comecemos por recordar a Representação de Números Reais em Vírgula Flutuante, também chamada **Notação Científica Normalizada de base *b***.

$$x \in \mathbb{R} \setminus \{0\} \quad \leftrightarrows \quad \pm 0.d_1 d_2 d_3 \ldots \times b^e$$

com $d_i \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$ e $e \in \mathbb{Z}$.

onde:

- **Base**: do sistema de numeração, habitualmente $b = 10$
- **Mantissa**: $d_1 d_2 d_3 \ldots$, é uma sequência de dígitos, possivelmente **infinita**.
- **Expoente**: $e \in \mathbb{Z}$
- **Representação Normalizada**: $d_1 \neq 0$. Garante a unicidade da representação.
- O **zero** não tem representação normalizada.

Mas como são representados os Números Reais num computador?

{ sob o ponto de vista do utilizador e não da representação interna }

- A representação dos números reais **não é exacta**.
  O número de dígitos da mantissa determina o grau de precisão.
- A grandeza dos números reais é **limitada**.
  O número de dígitos do expoente determina a grandeza máxima.
- A variação dos números reais representados é **discreta** e não contínua.
  A densidade de valores representados decresce exponencialmente com a grandeza dos números.

⇨ Representação **discreta** e **limitada** do conjunto $\mathbb{R}$.

Por isso, em vez de $\mathbb{R}$ utilizamos um subconjunto finito chamado $\mathbb{F}$ (*floating point*).

$$x \in \mathbb{R} \xrightarrow[\text{corte}]{\text{arredondamento}} fl(x) \in \mathbb{F}$$

Representação Normalizada dos elementos de $\mathbb{F}$ na base *b*.

$$fl(x) \in \mathbb{F}(b, t, q_1, q_2) \quad \leftrightarrows \quad (-1)^s \times 0.d_1\, d_2\, d_3 \ldots d_t \times b^e$$

com $s \in \{0, 1\}$, $d_i \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$ e $e \in (q_1, q_2)$, $q_1 < 0 < q_2$.

onde:

- **Base**: do sistema de numeração.
- **Sinal**: representado por $s \in \{0, 1\}$.
- **Mantissa**: $d_1 d_2 d_3 \ldots d_t$, é uma sequência **finita** de ***t*** dígitos.
- **Expoente**: $e \in (q_1, q_2)$ com $q_1 < 0$ e $q_2 > 0$ **finitos**.
- **Representação Normalizada**: $d_1 \neq 0$. Garante a unicidade da representação.
- **$0 \notin \mathbb{F}$**

Os valores exactos dos parâmetros de uma representação dependem do Processador Aritmético e da Linguagem de Programação utilizados.

Qual o erro provocado pela representação de $x \in \mathbb{R}$ por $fl(x) \in \mathbb{F}$?

$$x \in \mathbb{R} \xrightarrow[\text{corte}]{\text{arredondamento}} fl(x) \in \mathbb{F}$$

Seja,

$$fl(x) = (-1)^s \, . \, 0.d_1d_2d_3\ldots d_t \, . \, b^e \quad \in \mathbb{F}(b, t, q_1, q_2)$$

a representação de um dado $x \in \mathbb{R}$

$$x = (-1)^s \, . \, 0.d_1d_2d_3\ldots d_t \, d_{t+1} \ldots \, . \, b^e$$

onde assumimos que $e \in (q_1, q_2)$, ou seja, que o erro afecta apenas a representação da mantissa.

O ***erro absoluto*** cometido ***por corte*** será então:

$$\begin{aligned} \Delta_{corte} = |\, fl(x) - x \,| &= 0.000\ldots0d_{t+1} \ldots \, . \, b^e \\ &= 0.d_{t+1} \ldots \, . \, b^{e-t} \\ &\leq 0.9\ldots \, . \, b^{e-t} \\ &\leq b^{e-t} \end{aligned}$$

e o ***erro absoluto*** cometido ***por arredondamento simétrico*** será:

$$\begin{aligned} \Delta_{arred} = |\, fl(x) - x \,| &= 0.000\ldots0d_{t+1} \ldots \, . \, b^e \\ &\leq 0.5 \, . \, b^{e-t} = \tfrac{1}{2} \, b^{e-t} \end{aligned}$$

Procurando majorantes para os respectivos erros relativos,

teremos para o ***erro relativo*** cometido ***por corte***:

$$\begin{aligned} r_{corte} &= |\, fl(x) - x \,| \,/\, |\, x \,| \\ &\leq b^{e-t} \,/\, |\, x \,| \;=\; b^{e-t} \,/\, 0.d_1d_2 \ldots . \; b^e \\ &\leq b^{e-t} \,/\, 0.1 \;.\; b^e \;\; = b^{1-t} \end{aligned}$$

e para o ***erro relativo*** cometido ***por arredondamento simétrico***:

$$\begin{aligned} r_{arred} &= |\, fl(x) - x \,| \,/\, |\, x \,| \\ &\leq \tfrac{1}{2}\, b^{e-t} \,/\, |\, x \,| \;=\; \tfrac{1}{2}\, b^{e-t} \,/\, 0.d_1d_2 \ldots . \; b^e \\ &\leq \tfrac{1}{2}\, b^{e-t} \,/\, 0.1 \;.\; b^e \;\; = \tfrac{1}{2}\, b^{1-t} \end{aligned}$$

Estes resultados são referidos **no exercício 4 da 1ªfolha prática**:

4. Sendo $fl(x)$ a representação de x no sistema de vírgula flutuante $\mathbb{F}(b, t, q_1, q_2)$ (de base b, t dígitos na mantissa e pelo intervalo (q_1, q_2) de variação do expoente), mostre que, para $x \neq 0$

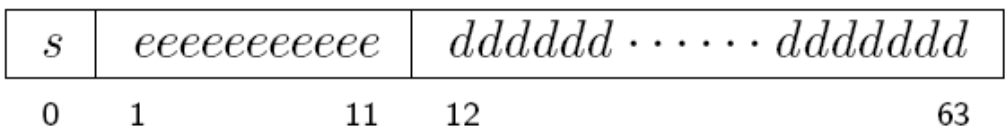
$$\frac{|fl(x) - x|}{|x|} \leq \begin{cases} b^{1-t}\,, & \text{se } fl(x) \text{ obtido por corte;} \\ \frac{1}{2}b^{1-t}\,, & \text{se } fl(x) \text{ obtido por arredondamento.} \end{cases}$$

A norma IEEE 754

(ver: http://www.cs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF *)*

- representação **binária** normalizada em vírgula flutuante
- **formato simples**: palavras de 32 bits
- **formato duplo**: palavras de 64 bits
- base de representação: ***b* = 2**
- por defeito, o MATLAB usa o **formato duplo**

Distribuição dos bits:



- $s \rightsquigarrow$ bit de sinal
- $e \rightsquigarrow$ bit do expoente
- $d \rightsquigarrow$ bit da mantissa

Se 1 ≤ expoente ≤ 2046

$$00000000001 \leq eeeeeeeeeee \leq 11111111110$$

então o valor V representado é

$$V = (-1)^s \times 2^{E-1023} \times (1.D)$$

$$V = (-1)^s \times 2^{E-1023} \times (1.D)$$

- (1.*D*) representa a mantissa normalizada (1 ≤ *mantissa* < 2).

  O primeiro bit da mantissa é sempre 1 (bit implícito) e não é armazenado.

- O expoente *E* é "enviesado".

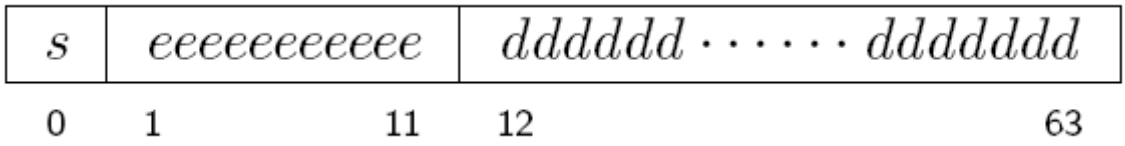
  Para permitir a representação de expoentes negativos:

$$2^{-1022} = 2^{1-1023} \leq 2^{E-1023} \leq 2^{2046-1023} = 2^{1023}$$

Assim, o menor número real positivo representável é 2^{-1022}.

Qualquer valor inferior iria gerar uma situação de *underflow*.

Representação do Zero



Se **expoente = 0** e **mantissa = 0**

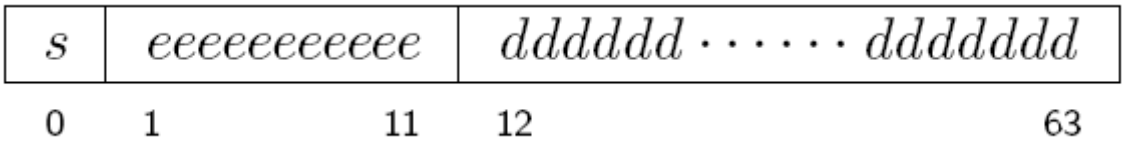
$$eeeeeeeeeee = 00000000000$$

$$ddddd \cdots\cdots ddddddd = 00000 \cdots\cdots 00000$$

então o valor *V* representado é: se ***s* = 0** então ***V* = +0**

se ***s* = 1** então ***V* = −0**

Representação da vizinhança de Zero (*underflow* gradual)



A vizinhança do zero é tratada de modo diferente, por forma a permitir uma representação ***mais densa*** dos números de pequena grandeza.

Se **expoente = 0** e **mantissa ≠ 0**

$$eeeeeeeeeee = 00000000000$$

$$ddddd \cdots\cdots ddddddd \neq 00000 \cdots\cdots 00000$$

então o valor V representado é:

$$V = (-1)^s \times 2^{-1022} \times (0.D)$$

- ($0.D$) representa uma mantissa não normalizada (0 < *mantissa* < 1).

  O primeiro bit da mantissa é sempre 0 (bit implícito) e não é armazenado.

- A técnica de *underflow* gradual permite representar uma mais vasta gama de valores na vizinhança de zero.

- O menor número positivo representável é agora,

$$2^{-1022} \times 0.0000...0001 = 2^{-1022} \times 2^{-52} = 2^{-1074}$$

- Números positivos inferiores são colocados a zero.

Caso dos Infinitos e NANs

s	$eeeeeeeeeee$	$dddddd \cdots\cdots dddddddd$
0	1 11	12 63

Se **expoente = 2047** e **mantissa = 0**

$$eeeeeeeeeee = 11111111111$$

$$dddddd \cdots\cdots ddddddd = 00000 \cdots\cdots 00000$$

então o valor *V* representado é: se ***s* = 0** então ***V* = +∞**

se ***s* = 1** então ***V* = −∞**

Se **expoente = 2047** e **mantissa ≠ 0**

então o valor *V* representado é:

V*= *NAN (Not A Number)

1.4. Erros de método ou de truncatura

Cometem-se **erros de truncatura** quando se usam:

- Métodos de discretização: aproximação de um problema de natureza contínua por outro de natureza discreta.

  Exemplos:

  - substituições de derivadas por razões incrementais.
  - integrais por somatórios
  - séries por somas de um número finito de termos

- Métodos iterativos: a partir de uma aproximação inicial, a solução é obtida (teoricamente) ao fim de um número infinito de operações.

  Na prática os processos iterativos são terminados ao fim de um número finito de operações.

Um exemplo:

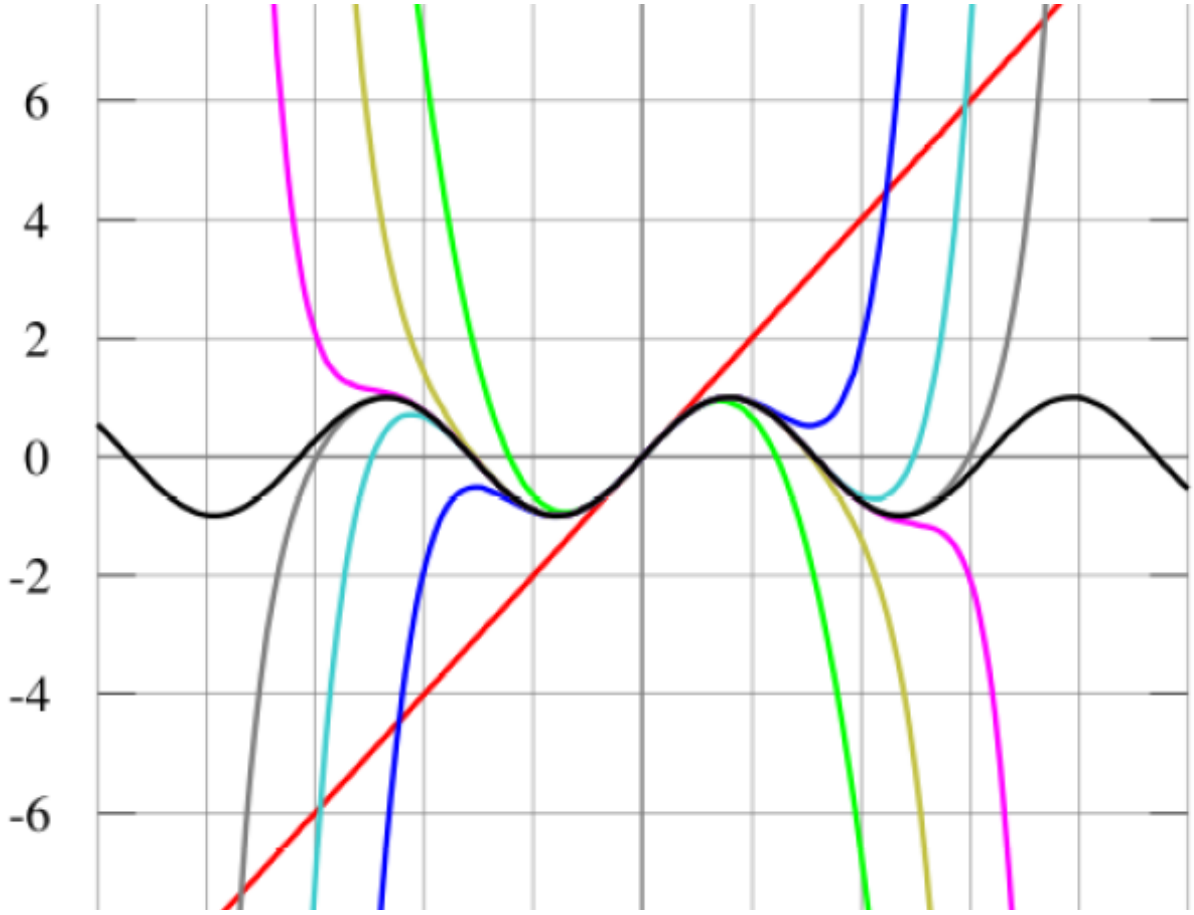
Consideremos o seguinte desenvolvimento em **série de Maclaurin** da função *seno*,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots$$

Dada a impossibilidade prática de calcular um número infinito de termos, só poderemos considerar somas parciais, que são os sucessivos **polinómios de Maclaurin**.

Cada polinómio constitui uma **aproximação** da função *seno* pretendida.

(*figura de*: http://en.wikipedia.org/wiki/Taylor_series)



Polinómios de Maclaurin:

x

$x - x^3 / 3!$

$x - x^3 / 3! + x^5 / 5!$

$x - x^3 / 3! + x^5 / 5! - x^7 / 7!$

$x - x^3 / 3! + x^5 / 5! - x^7 / 7! + x^9 / 9!$

...

Caso geral:

Teorema 1.8 (de Taylor) *Assuma que* $f \in C^{n+1}[a,b]$ *e seja* $x_0 \in [a,b]$. *Então, para todo o* $x \in (a,b)$, *existe um número* $c = c(x)$ *entre* x_0 *e* x *tal que*

$$f(x) = p_n(x) + R_n(x),$$

onde

$$p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

e

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)^{n+1}.$$

$p_n(x)$ é designado **o polinómio de Taylor de ordem n para** f **em torno de** x_0 e $R_n(x)$ é designado **o erro de truncatura associado a** $p_n(x)$.

No caso particular de $x_0 = 0$, o polinómio de Taylor é designado **polinómio de Maclaurin** e a série de Taylor é chamada **série de Maclaurin**.

Cálculo do erro de truncatura:

- $p_n(x)$ é uma aproximação de $f(x)$ com erro absoluto $|R_n(x)|$
- $R_n(x)$ não pode ser calculado porque se desconhece c
- **mas** é possível calcular um limite superior para $|R_n(x)|$,

  determinando um majorante para $|f^{(n+1)}(c)|$ com $c \in inter(x_0, x)$

voltando ao exemplo:

Analisemos a aproximação $p_7(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$

com erro de truncatura $R_7(x) = \frac{f^{(8)}(c)}{8!}x^8, \; c \in inter\ (0,x)$

Calculemos um majorante do erro cometido no ponto $\pi/4$:

Como $f^{(8)}(x) = \sin x$ e $\mid \sin(c) \mid \leq 1$ podemos estabelecer:

$$\mid R_7(\pi/4) \mid \leq \frac{(\pi/4)^8}{8!} \approx 0.359086 \times 10^{-5}$$

Caso particular das Séries Alternadas Convergentes:

$$S = \sum_{k=0}^{\infty} (-1)^k a_k, \quad (a_k > 0),$$

quando aproximamos S *por* $S_n = \sum_{k=0}^{n} (-1)^k a_k$, *o erro* $R_n = S - S_n$ *é um erro de truncatura. Um majorante para o valor absoluto do erro de truncatura pode ser tomado como o módulo do valor correspondente ao primeiro termo desprezado, isto é* $|R_n| = |S - S_n| \leq |a_{n+1}|$.

para o mesmo exemplo:

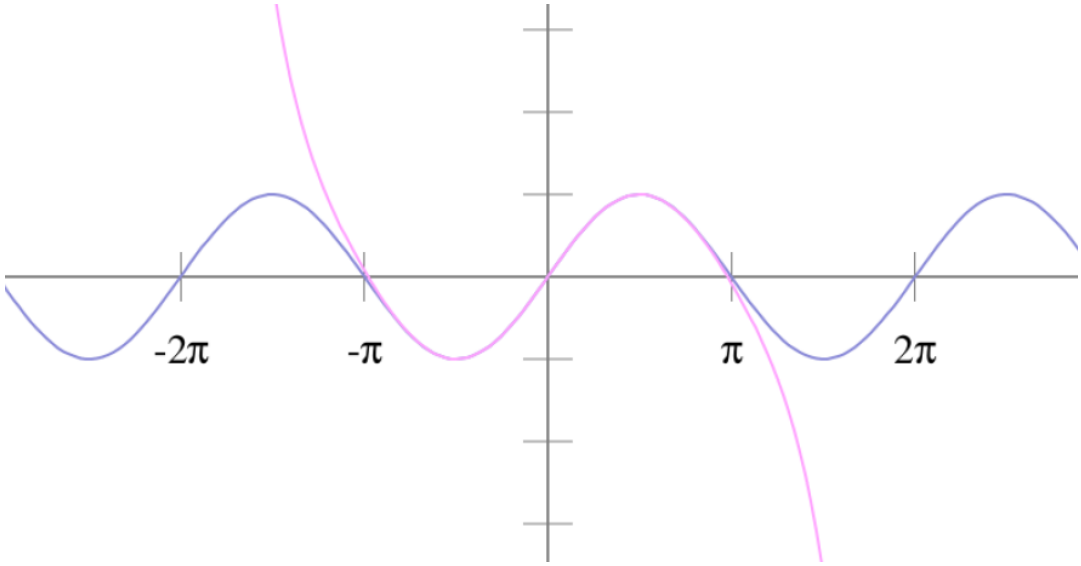
$$\mid R_7(\pi/4) \mid \leq \frac{(\pi/4)^9}{9!} \approx 0.313362 \times 10^{-6}$$

Observação:

Enquanto que a função seno é **periódica**, a aproximação polinomial já não o é.

Assim, o cálculo dos sucessivos valores de $\sin(\pi/4 \pm 2k\pi)$, $k = 0, 1, 2, \ldots$

virá afectado de um erro cada vez maior.



Podemos calcular esse efeito utilizando o majorante anterior:

k	$R_7(\pi/4 + 2k\pi)$
0	3.13362×10^{-7}
1	6.28319
2	12.5664
3	18.8496
4	25.1327
5	31.4159
6	37.6991
7	43.9823
8	50.2655
9	56.5487
10	62.8319

Por isso, no cálculo aproximado das funções trigonométricas comuns,

é **indispensável reduzir** o valor do ângulo ao intervalo $[-\pi, +\pi]$.

A questão inversa:

Dado um erro, quantos termos somar?

Exemplo: e = 2.71828182845904523536028747135266249775724709369995957 49 ...

Para calcular uma aproximação de ***e*** pelo desenvolvimento em **série de Maclaurin**

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

qual a ordem do menor polinómio que garante um erro inferior a 10^{-6} ?

Nota:

0	1.0000000000000
1	2.0000000000000
2	2.5000000000000
3	2.6666666666667
4	2.7083333333333
5	2.7166666666667
6	2.7180555555556
7	2.7182539682540
8	2.7182787698413
9	2.7182815255732
10	2.7182818011464
11	2.7182818261985
12	2.7182818282862

Calculemos o valor do erro de truncatura,

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}\,(x - x_0)^{n+1}$$

para $x = 1$, $x_0 = 0$, $f(x) = e^x$:

$$R_n(1) = \frac{e^c}{(n+1)!}$$

para $c \in inter(0, 1)$, um majorante aceitável poderá ser,

$$R_n(1) = \frac{e^c}{(n+1)!} < \frac{3}{(n+1)!}$$

Resta encontrar uma ordem n capaz de garantir que:

$$\frac{3}{(n+1)!} < 10^{-6}$$

ou seja, que:

$$(n+1)! > 3 \times 10^{+6}$$

e podemos verificar que este valor é atingido para $\mathbf{n \geq 9}$, tal como previsto.

(ver: http://en.wikipedia.org/wiki/Image:Exp_series.gif
Contém uma animação das sucessivas aproximações, bem como o respectivo programa em MATLAB)

Propagação de erros

Consideremos um determinado problema de cálculo numérico,

$$Y = F(X)$$

Mesmo que seja possível executar F de forma exacta, qualquer perturbação no valor dos dados irá afectar o valor dos resultados. São os **Erros Propagados:**

$$\bar{Y} = F(\bar{X})$$

Por outro lado, mesmo que os dados sejam exactos, o método de cálculo pode ser aproximado. Os resultados virão afectados de **Erros Gerados:**

$$\bar{Y} = \bar{F}(X)$$

Na maior parte das vezes, ocorrem sucessivas combinações desses dois tipos de erros:

$$\bar{Y} = \bar{F}(\bar{X})$$

Como se propagam os Erros?

por exemplo:

$$x = 2, \quad \bar{x} = 2.00005, \quad \Delta_{\bar{x}} = 5 \times 10^{-5}, \quad r_{\bar{x}} = 2.5 \times 10^{-5}$$

$$y = f(x) = \frac{x}{20} = 0.1, \quad \bar{y} = 0.100003, \quad \Delta_{\bar{y}} = 2.5 \times 10^{-6}, \quad r_{\bar{y}} = 2.5 \times 10^{-5}$$

$$y = f(x) = \frac{20}{x} = 10, \quad \bar{y} = 9.99975, \quad \Delta_{\bar{y}} = 2.49994 \times 10^{-4}, \quad r_{\bar{y}} = 2.49994 \times 10^{-5}$$

$$y = f(x) = \frac{20}{x^2} = 5, \quad \bar{y} = 4.99975, \quad \Delta_{\bar{y}} = 2.49991 \times 10^{-4}, \quad r_{\bar{y}} = 4.99981 \times 10^{-5}$$

porquê?

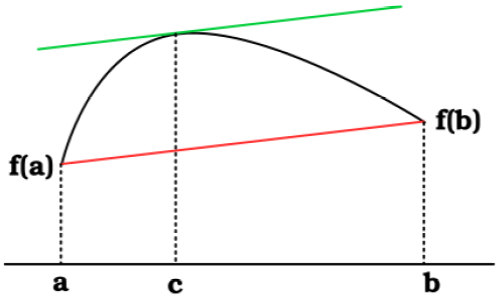
Procuremos uma fórmula geral para a propagação dos erros:

Mas antes disso recordemos:

Teorema do Valor Médio (Lagrange)

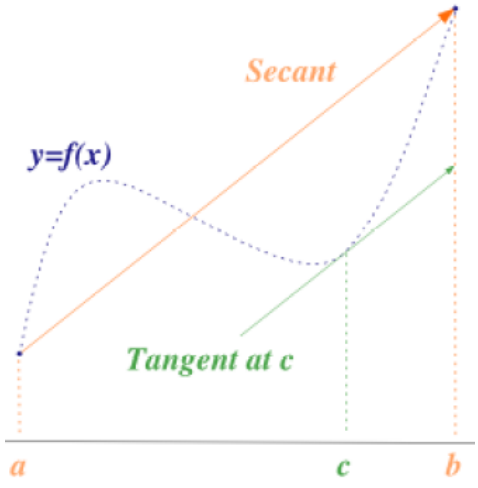
Seja $f : [a, b] \longrightarrow \mathbb{R}$ contínua em $[a, b]$ e derivável em (a, b) então existe $c \in (a, b)$ tal que

$$f(b) - f(a) = f'(c)(b - a) \quad .$$



Portanto, existe (pelo menos) um ponto **c** onde: $f'(c) = \dfrac{f(b) - f(a)}{b - a}$

ou seja, onde a ***tangente*** é paralela à ***secante***:



Exercício + Corolário do T.V.M.:

Seja $y = f(x)$, onde f é uma função continuamente diferenciável em $\mathbb{R}$. Admita que $\bar{y} = f(\bar{x})$, isto é, $\bar{y}$ é obtido usando aritmética exacta com dados ligeiramente perturbados $(\bar{x})$.

(a) Utilizando o teorema do valor médio, mostre que $\Delta_{\bar{y}} = |f'(\xi)|\,\Delta_{\bar{x}}$, $\xi \in (x, \bar{x})$, onde $\Delta_{\bar{y}}$ e $\Delta_{\bar{x}}$ são, respectivamente, o erro absoluto de $\bar{y}$ e o erro absoluto de $\bar{x}$.

(b) A partir da igualdade anterior, prove que $r_{\bar{y}} \approx \dfrac{|x\, f'(x)|}{|f(x)|}\, r_{\bar{x}}$, onde $r_{\bar{y}}$ e $r_{\bar{x}}$ são, respectivamente, o erro relativo de $\bar{y}$ e o erro relativo de $\bar{x}$. Comente este resultado, interpretando o seu significado.
(normalmente $\frac{|x\, f'(x)|}{|f(x)|}$ é designado por **número de condição** de f em x e nota-se por $cond\ f(x)$)

Resolução + Demonstração + Comentários:

Sejam $y = f(x)$, $f \in C^1(\mathbb{R})$, $\bar{y} = f(\bar{x})$, $\bar{x}$ uma aproximação de x

(a)

$$\begin{aligned}
\Delta\bar{y} &= |y - \bar{y}| \quad \text{por def. de erro absoluto} \\
&= |f(x) - f(\bar{x})| \quad \text{por hipótese} \\
&= |f'(\xi)|\,|x - \bar{x}| \quad \text{pelo T.V.M., para algum } \xi \in int(x, \bar{x}) \\
&= |f'(\xi)|\,\Delta_{\bar{x}} \quad \text{por def. de erro absoluto}
\end{aligned}$$

$$\Delta\bar{y} = |f'(\xi)|\,\Delta_{\bar{x}}$$

(b)

$$r_{\bar{y}} = \frac{\Delta_{\bar{y}}}{|y|} \quad \text{por def. de erro relativo}$$

$$= \frac{|f'(\xi)|\ |x-\bar{x}|}{|y|} \quad \text{pela alínea anterior}$$

$$= \frac{|x|\ |f'(\xi)|}{|y|}\ \frac{|x-\bar{x}|}{|x|}$$

$$= \frac{|x|\ |f'(\xi)|}{|y|}\ r_{\bar{x}} \quad \text{por def. de erro relativo}$$

$$\approx \frac{|x|\ |f'(x)|}{|f(x)|}\ r_{\bar{x}} \quad \text{porque } x \approx \bar{x} \text{ e } \xi \in int(x,\bar{x})$$

$$r_{\bar{y}} \approx \frac{|x\ \ f'(x)|}{|f(x)|}\ r_{\bar{x}}$$

cond f(x) = $\dfrac{|x\ \ f'(x)|}{|f(x)|}$ **número de condição** de f em x

cond f(x) é um indicador do efeito da propagação do erro relativo, no valor da função ***f*** no ponto ***x***, que nos permite avaliar em que condições a função é ***bem*** ou ***mal condicionada***.

exemplo:

Analisemos os efeitos da propagação de erros nas funções x^n e n^x com $n \in \mathbb{N}$

$$f(x) = x^n \text{ com } n \in \mathbb{N}$$

$$cond\ f(x) = \frac{|x\ f'(x)|}{|f(x)|} = \frac{|x\ n\ x^{n-1}|}{|x^n|} = n$$

Verificamos que a propagação do erro relativo depende apenas de n e não de x.

Com efeito para,

$$x = 2, \quad \bar{x} = 2.00005, \quad \Delta_{\bar{x}} = 5 \times 10^{-5}, \quad r_{\bar{x}} = 2.5 \times 10^{-5}$$

obtemos:

			$\Delta_{\bar{y}}$	$r_{\bar{y}}$
2	4	4.0002	2.00002×10^{-4}	5.00006×10^{-5}
6	64	64.0096	9.6006×10^{-3}	1.50009×10^{-4}
10	1024	1024.26	2.56029×10^{-1}	2.50028×10^{-4}
14	16384	16389.7	5.73533	3.50057×10^{-4}
18	262144	262262.	1.1799×10^{2}	4.50096×10^{-4}
22	4194304	4.19661×10^{6}	2.30747×10^{3}	5.50144×10^{-4}

e para,

$$x = 20, \quad \bar{x} = 20.0005, \quad \Delta_{\bar{x}} = 5 \times 10^{-4}, \quad r_{\bar{x}} = 2.5 \times 10^{-5}$$

obtemos:

400.02	2.00002×10^{-2}	5.00006×10^{-5}
6.40096×10^{7}	9.6006×10^{3}	1.50009×10^{-4}
1.02426×10^{13}	2.56029×10^{9}	2.50028×10^{-4}
1.63897×10^{18}	5.73533×10^{14}	3.50057×10^{-4}
2.62262×10^{23}	1.1799×10^{20}	4.50096×10^{-4}
4.19661×10^{28}	2.30747×10^{25}	5.50144×10^{-4}

$$f(x) = n^x \text{ com } n \in \mathbb{N}$$

$$cond\ f(x) = \frac{|x\ f'(x)|}{|f(x)|} = \frac{|x\ n^x\ \ln n|}{|n^x|} = |x\ \ln n|$$

Neste caso, a propagação do erro relativo depende de ***x*** mais do que de ***n***

Com efeito para,

$$x = 2, \quad \bar{x} = 2.00005, \quad \Delta_{\bar{x}} = 5 \times 10^{-5}, \quad r_{\bar{x}} = 2.5 \times 10^{-5}$$

obtemos:

			$\Delta_{\bar{y}}$	$r_{\bar{y}}$
2	4	4.00014	1.38632×10^{-4}	3.4658×10^{-5}
6	36	36.0032	3.22531×10^{-3}	8.9592×10^{-5}
10	100	100.012	1.15136×10^{-2}	1.15136×10^{-4}
14	196	196.026	2.58645×10^{-2}	1.31962×10^{-4}
18	324	324.047	4.68274×10^{-2}	1.44529×10^{-4}
22	484	484.075	7.4809×10^{-2}	1.54564×10^{-4}

e para,

$$x = 20, \quad \bar{x} = 20.0005, \quad \Delta_{\bar{x}} = 5 \times 10^{-4}, \quad r_{\bar{x}} = 2.5 \times 10^{-5}$$

obtemos:

1.04894×10^{6}	3.63472×10^{2}	3.46634×10^{-4}
3.65944×10^{15}	3.27695×10^{12}	8.96281×10^{-4}
1.00115×10^{20}	1.15196×10^{17}	1.15196×10^{-3}
8.37787×10^{22}	1.10476×10^{20}	1.3204×10^{-3}
1.27667×10^{25}	1.84369×10^{22}	1.44623×10^{-3}
7.06521×10^{26}	1.0911×10^{24}	1.54672×10^{-3}

Como se propagam os erros nas operações aritméticas?

No que se segue, $\overline{p}$ e $\overline{q}$ são valores aproximados dos números *p* e *q*, ambos com o mesmo sinal e são desprezados os erros de arredondamento das próprias operações.

Adição:

Erro Absoluto:

$$\begin{aligned}\Delta_{\overline{p}+\overline{q}} &= |(p+q)-(\overline{p}+\overline{q})| = |(p-\overline{p})+(q-\overline{q})| \\ &\leq |p-\overline{p}|+|q-\overline{q}| = \Delta_{\overline{p}}+\Delta_{\overline{q}}\end{aligned}$$

O erro absoluto da soma de dois números é limitado pela soma dos erros absolutos individuais.

Erro Relativo:

$$\begin{aligned}r_{\overline{p}+\overline{q}} &= \frac{\Delta_{\overline{p}+\overline{q}}}{|p+q|} \leq \frac{\Delta_{\overline{p}}+\Delta_{\overline{q}}}{|p+q|} = \frac{\Delta_{\overline{p}}}{|p|}\frac{|p|}{|p+q|}+\frac{\Delta_{\overline{q}}}{|q|}\frac{|q|}{|p+q|} \\ &= r_{\overline{p}}\frac{|p|}{|p+q|}+r_{\overline{q}}\frac{|q|}{|p+q|}\end{aligned}$$

Considerando, $r = \max\left\{r_{\overline{p}}, r_{\overline{q}}\right\}$ conclui-se que $r_{\overline{p}+\overline{q}} \leq r$

O erro relativo da soma de dois números é limitado pelo maior dos erros relativos individuais.

Subtracção:

Erro Absoluto:

$$\Delta_{\overline{p}-\overline{q}} = |(p-q)-(\overline{p}-\overline{q})| = |(p-\overline{p})-(q-\overline{q})|$$
$$\leq |p-\overline{p}| + |q-\overline{q}| = \Delta_{\overline{p}} + \Delta_{\overline{q}}$$

O erro absoluto da subtracção de dois números é limitado pela soma dos erros absolutos individuais.

Erro Relativo:

$$r_{\overline{p}-\overline{q}} = \frac{\Delta_{\overline{p}-\overline{q}}}{|p-q|} \leq \frac{\Delta_{\overline{p}}+\Delta_{\overline{q}}}{|p-q|} = \frac{\Delta_{\overline{p}}}{|p|}\frac{|p|}{|p-q|} + \frac{\Delta_{\overline{q}}}{|q|}\frac{|q|}{|p-q|}$$
$$= r_{\overline{p}}\frac{|p|}{|p-q|} + r_{\overline{q}}\frac{|q|}{|p-q|}$$

Fenómeno de **Cancelamento Subtractivo:**

Quando se subtraem quantidades muito próximas (diferença $p - q$ pequena) o erro relativo pode vir muito elevado.

Multiplicação e Divisão:

Uma estimativa para o erro relativo no produto (divisão) é dada pela soma dos erros relativos dos operandos (desde que estes venham afectados por um erro relativo pequeno).

{ Demonstre ... }

1.5. Transformação de Fórmulas

exemplo1:

Calcular $e^{-100} = 3.7200759760208359630 \times 10^{-44}$ pelo desenvolvimento,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

Usando, por exemplo, o seguinte programa em MATLAB com $x = -100$,

```
soma = 1;
termo = x;
n = 1;
while  abs(termo) > 10^(-50)
       soma = soma + termo;
       n = n + 1;
       termo = termo * x / n;
end
disp(['soma de ' int2str(n) ' termos = ' num2str(soma) ])
disp(['proximo termo = ' num2str(termo) ])
disp(['valor pretendido = ' num2str( exp(x) ) ])
```

verificamos que é tarefa praticamente impossível. O que acontece? Porquê?

Como a grandeza dos termos tende para zero (série convergente), a partir de certa ordem irão ocorrer sucessivos **cancelamentos subtractivos** entre números muito pequenos de sinal alternado.

Para resolver este problema, basta constatar que $e^{-x} = 1 / e^{x}$ e executar o mesmo programa para $x = 100$, invertendo o resultado obtido.

exemplo2:

Calcular $\sqrt{x+1}-\sqrt{x}$ para valores grandes de *x*.

Tente no MATLAB e verificará que, por exemplo, para números de grandeza 10^{20} o resultado virá (erradamente) nulo.

Uma transformação adequada da fórmula poderá ser,

$$\left(\sqrt{x+1}-\sqrt{x}\right)\frac{\sqrt{x+1}+\sqrt{x}}{\sqrt{x+1}+\sqrt{x}}=\frac{1}{\sqrt{x+1}+\sqrt{x}}$$

que já permitirá obter resultados razoáveis.

exemplo3:

Calcular $\dfrac{\sin x}{x}$ para valores muito pequenos de *x*.

A fim de evitar a divisão por uma quantidade muito pequena, é preferível o desenvolvimento,

$$\frac{\sin x}{x}=1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots+(-1)^{n-1}\frac{x^{2n-2}}{(2n-1)!}+\ldots$$

1.6. Condicionamento e estabilidade

Num **problema** *P* existem **dados de entrada**, que podemos agrupar muito geralmente num vector *x* e resultados (**dados de saída**), que podemos designar por *y*

$$y = P(x)$$

Definição:

Um problema diz-se **bem condicionado** (ou matematicamente **estável**) se pequenos erros relativos nos dados produzem pequenos erros relativos no resultado.

Caso contrário, diz-se **mal condicionado** (ou matematicamente **instável**).

Exemplo de problema mal condicionado:

Resolver a equação $x^2 - \frac{x}{3} + \frac{1}{36} = 0$

que tem raízes reais $x_1 = x_2 = \frac{1}{6}$

Uma pequena variação nos valores de *1/3* e de *1/36*, por exemplo causada por arredondamentos a 6 casas decimais, resulta na equação:

$$x^2 - 0.333333x + 0.027778 = 0$$

que não tem raízes reais!

Na resolução de um problema P por utilização de um algoritmo A, para além dos erros dos dados temos de considerar os **erros de arredondamento** que se irão propagar ao longo da execução do algoritmo. Assim, considerando os dados de entrada x e os resultados y

$$y = A(x)$$

Definição:

Um **método** (ou algoritmo) diz-se **computacionalmente** (ou numericamente) **estável** se a acumulação e propagação dos erros de arredondamento provoca um pequeno erro relativo no resultado.

Caso contrário, diz-se **computacionalmente** (ou numericamente) **instável**.

nota:

Nenhum algoritmo, quando aplicado a um problema mal condicionado, poderá ser computacionalmente estável!