

MATEMÁTICA COMPUTACIONAL

Cap. 1. Representação de Números e Teoria de Erros

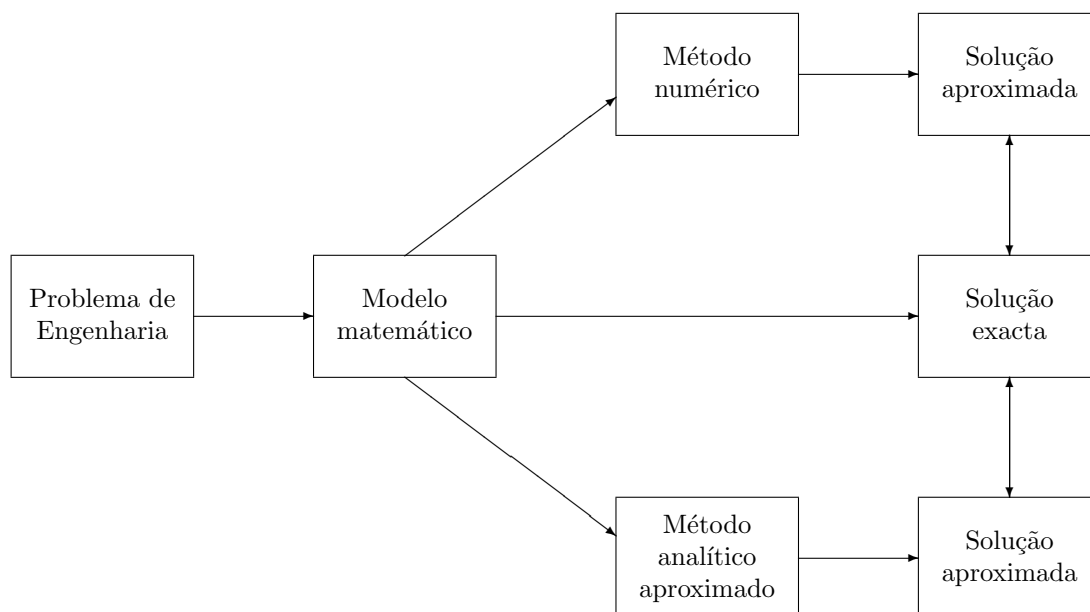
Filipe J. Romeiras

Departamento de Matemática
Instituto Superior Técnico

Setembro de 2007

Apontamentos das aulas da disciplina do mesmo nome do 2^o ano do
Mestrado Integrado em Engenharia Física Tecnológica

1. REPRESENTAÇÃO DE NÚMEROS E TEORIA DE ERROS



Tipos de erro

- **Erros inerentes:**

- ◇ modelo matemático incompleto;
- ◇ erros nos dados, parâmetros e constantes matemáticas;
- ◇ exemplo: corpo em movimento vertical na atmosfera terrestre sujeito à força de atracção gravitacional da Terra e à força de resistência do ar.

- **Erros de método (ou de truncatura ou de discretização):**

- ◇ exemplos:
 - cálculo do valor aproximado da raiz de uma função pelo método de Newton;
 - cálculo do valor aproximado de um integral usando a regra dos trapézios composta.

- **Erros computacionais:**

- ◇ erros de arredondamento;
- ◇ erros de “underflow” e “overflow”.

- **Erros de programação.**

Erro, erro absoluto, erro relativo ($\tilde{x} \approx x \in \mathbb{R}$):

Definição. Seja $x \in \mathbb{R}$ o valor exacto de uma grandeza real e \tilde{x} um valor aproximado de x . Definem-se:

Erro de \tilde{x} em relação a x : $e_{\tilde{x}} = x - \tilde{x}$

Erro absoluto de \tilde{x} em relação a x : $|e_{\tilde{x}}|$

Erro relativo de \tilde{x} em relação a $x \neq 0$: $\delta_{\tilde{x}} = \frac{e_{\tilde{x}}}{x}$, ou $|\delta_{\tilde{x}}|$

(Percentagem de erro: $100\delta_{\tilde{x}}(\%)$ ou $100|\delta_{\tilde{x}}|(\%)$)

Nota. O erro relativo é invariante numa mudança de escala, i.e., sendo $y = kx$, $\tilde{y} = k\tilde{x}$, onde k é uma constante não nula, então $\delta_{\tilde{y}} = \delta_{\tilde{x}}$.

Exemplo.

$$x = \frac{1}{3}, \quad \tilde{x} = 0.3333, \quad y = \frac{1}{3000}, \quad \tilde{y} = 0.0003.$$

$$e_{\tilde{x}} = e_{\tilde{y}} = 0.0000333\dots, \quad \delta_{\tilde{x}} \approx 0.0001(0.01\%) \quad \delta_{\tilde{y}} \approx 0.1(10\%)$$

Representação de números

• Um número inteiro $x \in \mathbb{Z} \setminus \{0\}$ é representado numa **base** $\beta \in \mathbb{N} \setminus \{0, 1\}$ (por exemplo, 10, 2, 16, 8, 12, 20, 60) por

$$x = \sigma(d_m\beta^m + d_{m-1}\beta^{m-1} + \dots + d_1\beta^1 + d_0\beta^0)$$

onde

$$\sigma \in \{+, -\}, \quad d_i \in \{0, 1, \dots, \beta - 1\}, \quad i = 0, 1, \dots, m, \quad d_m \neq 0$$

Simbolicamente escreve-se

$$x = \sigma(d_m d_{m-1} \dots d_1 d_0)_\beta$$

Exemplo. $198 = (198)_{10} = (11000110)_2$

Nota. Os números inteiros são representados exactamente num computador. Se neste reservarmos $N + 1$ bits para números inteiros podemos representar todos os números inteiros tais que

$$-(2^N - 1) \leq x \leq (2^N - 1)$$

Exemplo. $N = 31 : 2^N - 1 = 2.147.483.647$

• Um número real $x \in \mathbb{R} \setminus \{0\}$ é representado numa base $\beta \in \mathbb{N} \setminus \{0, 1\}$, por

$$x = \sigma(d_m\beta^m + d_{m-1}\beta^{m-1} + \dots + d_1\beta^1 + d_0\beta^0 +$$

$$+ d_{-1}\beta^{-1} + d_{-2}\beta^{-2} + \dots + d_{-n}\beta^{-n} + \dots)$$

onde

$$\sigma \in \{+, -\}, \quad d_i \in \{0, 1, \dots, \beta - 1\}, \quad i = m, m-1, \dots, \quad d_m \neq 0$$

Simbolicamente escreve-se

$$x = \sigma(d_m d_{m-1} \dots d_1 d_0 . d_{-1} d_{-2} \dots)_\beta$$

ou

$$x = \sigma(0.d_m d_{m-1} \dots d_1 d_0 d_{-1} d_{-2} \dots)_\beta \times \beta^{m+1}$$

Exemplo.

$$\begin{aligned} 19 &= (10011)_2, & 0.875 &= (0.111)_2, \\ 19.875 &= (10011.111)_2 = (0.10011111)_2 \times 2^{(101)_2} \\ 0.1 &= (0.1)_{10} = (0.11001100\dots)_2 \times 2^{-(11)_2} \end{aligned}$$

• Notação científica

$$x = \sigma m \beta^t$$

$$\text{(base)} \beta \in \mathbb{N} \setminus \{0, 1\}, \quad \text{(sinal)} \sigma \in \{+, -\}, \quad \text{(expoente)} t \in \mathbb{Z}$$

$$\text{(mantissa)} m = (0.a_1 a_2 \dots)_\beta \in [\beta^{-1}, 1[, \quad a_i \in \{0, 1, \dots, \beta - 1\}, \quad a_1 \neq 0$$

Definição. Sejam $\beta \in \mathbb{N} \setminus \{0, 1\}$, $n \in \mathbb{N} \setminus \{0\}$, $t^-, t^+ \in \mathbb{Z}$. Designa-se por **sistema de ponto flutuante** na base β , com n dígitos na mantissa, e expoentes variando entre t^- e t^+ , ao subconjunto dos números racionais

$$\mathbb{F} = \text{FP}(\beta, n, t^-, t^+) = \{x \in \mathbb{Q} : x = \sigma m \beta^t\} \cup \{0\}$$

$$\sigma \in \{+, -\}, \quad t \in \mathbb{Z}, \quad t^- \leq t \leq t^+,$$

$$m = (0.a_1 a_2 \dots a_n)_\beta \in [\beta^{-1}, 1 - \beta^{-n}], \quad a_i \in \{0, 1, \dots, \beta - 1\}, \quad a_1 \neq 0$$

Quando apenas for importante referir a base e o número de dígitos da mantissa, usamos a notação $\text{FP}(\beta, n)$.

Proposição.

$$\text{card}(\text{FP}(\beta, n, t^-, t^+)) = 2N + 1, \quad N = (t^+ - t^- + 1)(\beta - 1)\beta^{n-1}.$$

Nota. N é o número de racionais positivos de \mathbb{F} compreendidos entre

$$L^- = \beta^{-1} \times \beta^{t^-}, \quad L^+ = (1 - \beta^{-n}) \times \beta^{t^+}.$$

Os restantes elementos de \mathbb{F} são os simétricos destes e o número zero.

Exemplo: $\text{FP}(2, 3, -1, 1)$.

Exemplo: Sistemas de ponto flutuante definidos pela Norma IEC559 (International Electronic Commission, 1989).

Formato simples: $\text{FP}(2, 24, -125, 128)$

Formato duplo: $\text{FP}(2, 53, -1021, 1024)$

Arredondamentos

Questão. Dado um número $x \in \mathbb{R}_{\mathbb{F}}$ e $x \notin \mathbb{F}$ qual o número $\text{fl}(x) \in \mathbb{F}$ que o representa, onde $\mathbb{R}_{\mathbb{F}} = [-L^+, -L^-] \cup \{0\} \cup [L^-, L^+]$?

Definição. Dado $\mathbb{F} = \text{FP}(\beta, n, t^-, t^+)$, e sendo

$$x = \sigma(0.a_1a_2 \dots a_n a_{n+1} \dots)_\beta \times \beta^t,$$

definem-se as duas seguintes funções de arredondamento $\text{fl}_c : \mathbb{R}_{\mathbb{F}} \rightarrow \mathbb{F}$ e $\text{fl}_s : \mathbb{R}_{\mathbb{F}} \rightarrow \mathbb{F}$:

(i) Arredondamento por corte:

$$\text{fl}_c(x) = \sigma(0.a_1a_2 \dots a_n)_\beta \times \beta^t$$

(ii) Arredondamento simétrico (β par):

$$\text{fl}_s(x) = \begin{cases} \sigma(0.a_1a_2 \dots a_n)_\beta \times \beta^t, & 0 \leq a_{n+1} < \frac{\beta}{2} \\ \sigma[(0.a_1a_2 \dots a_n)_\beta + \beta^{-n}] \times \beta^t, & \frac{\beta}{2} \leq a_{n+1} < \beta \end{cases}$$

ou, de uma forma equivalente,

$$\text{fl}_s(x) = \text{fl}_c\left(x + \frac{1}{2}\beta^{t-n}\right)$$

Nota.

(i) $\text{fl}_c(x)$ é o número de \mathbb{F} mais perto de x entre 0 e x .

(ii) $\text{fl}_s(x)$ é o número de \mathbb{F} mais perto de x . Se houver dois números de \mathbb{F} igualmente perto de x então $\text{fl}_s(x)$ é o maior deles.

Exemplo. Sendo $\pi = 3.1415926535 \dots$ e $\mathbb{F} = \text{FP}(10, 7)$ então,

$$\text{fl}_c(\pi) = 0.3141592 \times 10^{+1}, \quad \text{fl}_s(\pi) = 0.3141593 \times 10^{+1}.$$

Exemplo. Sendo $0.1 = x = (0.11001100 \dots)_2 \times 2^{-3}$ e $\mathbb{F} = \text{FP}(2, 4)$ então

$$\text{fl}_c(x) = (0.1100)_2 \times 2^{-3} = \frac{3}{32} = (0.09375)_{10}$$

$$\text{fl}_s(x) = (0.1101)_2 \times 2^{-3} = \frac{13}{128} = (0.1015625)_{10}$$

Erros de arredondamento

Proposição. Sendo $x \in \mathbb{R}_{\mathbb{F}}$ e $\tilde{x} = \text{fl}(x) \in \mathbb{F} = \text{FP}(\beta, n, t^-, t^+)$, então:

(i) Arredondamento por corte:

$$|e_{\tilde{x}}| < \beta^{t-n}, \quad |\delta_{\tilde{x}}| < \beta^{1-n};$$

(ii) Arredondamento simétrico:

$$|e_{\tilde{x}}| < \frac{1}{2} \beta^{t-n}, \quad |\delta_{\tilde{x}}| < \frac{1}{2} \beta^{1-n}.$$

Dem.:

$$x = \sigma(0.a_1a_2 \dots a_n a_{n+1} \dots)_\beta \times \beta^t, \quad |x| \geq \beta^{-1+t}$$

(i) Arredondamento por corte:

$$\tilde{x} = \text{fl}_c(x) = \sigma(0.a_1a_2 \dots a_n)_\beta \times \beta^t$$

$$e_{\tilde{x}} = x - \tilde{x} = \sigma(0.0 \dots 0a_{n+1} \dots)_\beta \times \beta^t = \sigma(0.a_{n+1}a_{n+2} \dots)_\beta \times \beta^{t-n}$$

$$|e_{\tilde{x}}| < \beta^{t-n}$$

$$|\delta_{\tilde{x}}| = \frac{|e_{\tilde{x}}|}{|x|} < \beta^{1-n}$$

(ii) Arredondamento simétrico:

$$(a) 0 \leq a_{n+1} < \frac{\beta}{2}$$

$$(b) \frac{\beta}{2} \leq a_{n+1} < \beta$$

$$\tilde{x} = \text{fl}_s(x) = \sigma(0.a_1a_2 \dots a_n)_\beta \times \beta^t$$

$$\tilde{x} = \text{fl}_s(x) = \sigma[(0.a_1a_2 \dots a_n)_\beta + \beta^{-n}] \times \beta^t$$

$$e_{\tilde{x}} = \sigma(0.a_{n+1}a_{n+2} \dots)_\beta \times \beta^{t-n}$$

$$e_{\tilde{x}} = \sigma[(0.a_{n+1}a_{n+2} \dots)_\beta - 1] \times \beta^{t-n}$$

$$|e_{\tilde{x}}| < \frac{\beta}{2} \beta^{-1} \beta^{t-n} = \frac{1}{2} \beta^{t-n}$$

$$|e_{\tilde{x}}| \leq \frac{1}{2} \beta^{t-n}$$

$$|\delta_{\tilde{x}}| < \frac{1}{2} \beta^{1-n}$$

$$|\delta_{\tilde{x}}| < \frac{1}{2} \beta^{1-n}$$

Nota. O majorante do erro relativo depende de β , de n e do tipo de arredondamento, mas não depende de x .

Definição. Dado um sistema $\text{FP}(\beta, n, t^-, t^+)$ define-se a **unidade de arredondamento do sistema** por

$$u_c = \beta^{1-n}, \quad u_s = \frac{1}{2} \beta^{1-n}.$$

Exemplo. Sistemas definidos pela Norma IEC559.

$$\mathbb{F} = \text{FP}(2, 24, -125, 128): \quad u_s = 2^{-24} \approx 0.596046 \times 10^{-7}$$

$$\mathbb{F} = \text{FP}(2, 53, -1021, 1024): \quad u_s = 2^{-53} \approx 0.111022 \times 10^{-15}$$

Operações aritméticas num sistema de ponto flutuante

Definição. Sendo $\circ : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ uma operação aritmética, $\circ = +, -, \times, \div$, define-se a operação aritmética correspondente num sistema de ponto flutuante \mathbb{F} , $\boxed{\circ} : \mathbb{R}_{\mathbb{F}} \times \mathbb{R}_{\mathbb{F}} \rightarrow \mathbb{F}$, por

$$x \boxed{\circ} y = \text{fl}(\text{fl}(x) \circ \text{fl}(y))$$

Exemplo. Sendo $x = \pi$ e $y = \frac{333}{106}$ e considerando um sistema de ponto flutuante FP(10, 6, -10, 10) com arredondamento simétrico obtém-se:

$$\begin{aligned} x \boxed{+} y &= 0.628310 \times 10, & x \boxed{-} y &= 0.800000 \times 10^{-4}, \\ x \boxed{\times} y &= 0.986934 \times 10, & x \boxed{\div} y &= 0.100003 \times 10 \end{aligned}$$

Algarismos significativos

Definição. Seja

$$\tilde{x} = \sigma(0.a_1a_2 \dots a_n)_{10} \times 10^t$$

uma aproximação de x pertencente a FP(10, n). Diz-se que o algarismo a_i é **algarismo significativo** de \tilde{x} se

$$|e_{\tilde{x}}| \leq \frac{1}{2} 10^{t-i}.$$

Nota.

- (i) Se a_i , $i \geq 2$, é significativo então a_j , $1 \leq j < i$, são significativos.
- (ii) Se a_n é significativo então os n algarismos de \tilde{x} são significativos.

Nota. Se \tilde{x} é obtido de x por arredondamento simétrico então os n algarismos de \tilde{x} são significativos.

Exemplo. Sendo $\pi = 3.14159265358 \dots$, então:

- (i) a aproximação $\tilde{\pi} = 3.141592$ tem 6 algarismos significativos;

$$e_{\tilde{\pi}} \approx 0.6535 \times 10^{-6}, \quad \frac{1}{2} 10^{1-7} < |e_{\tilde{\pi}}| < \frac{1}{2} 10^{1-6}$$

- (ii) a aproximação $\tilde{\tilde{\pi}} = 3.141593$ tem 7 algarismos significativos.

$$e_{\tilde{\tilde{\pi}}} \approx -0.3465 \times 10^{-6}, \quad \frac{1}{2} 10^{1-8} < |e_{\tilde{\tilde{\pi}}}| < \frac{1}{2} 10^{1-7}$$

Exemplo. Número de algarismos significativos da representação de um número real nos sistemas de ponto flutuante definidos pela Norma IEC559 (considerando arredondamento simétrico).

Formato simples: FP(2, 24, -125, 128), $u \approx 0.596046 \times 10^{-7}$
6 ou 7 algarismos significativos

Formato duplo: FP(2, 53, -1021, 1024), $u \approx 0.111022 \times 10^{-15}$
15 ou 16 algarismos significativos

$$x = \sigma m 10^t, \quad x - \tilde{x} = x \delta_{\tilde{x}}, \quad |\delta_{\tilde{x}}| < u = m_u 10^{t_u}$$

$$|x - \tilde{x}| < m m_u 10^{t+t_u}, \quad 0.1 m_u \leq m m_u < m_u$$

Erros de “overflow” e “underflow”

Definição. Os erros de “overflow” e “underflow” ocorrem quando $t \notin \{t^-, \dots, t^+\}$. Se $t > t^+$ tem-se “overflow” enquanto se $t < t^-$ tem-se “underflow”.

Exemplo. Cálculo de $|x + iy|$ para $x = y = 10^{20}$ no sistema FP(2, 24, -125, 128).

Exemplo. Cálculo das raízes de $ax^2 + bx + c = 0$ para $a = 10^{20}$, $b = -3 \times 10^{20}$, $c = 2 \times 10^{20}$ no sistema FP(2, 24, -125, 128).

Propagação de erros (teoria linearizada)

Definição. Diz-se que

$$f(x) \doteq h(x), \quad \text{quando } x \rightarrow x^*,$$

i.e., f é igual a h em primeira aproximação quando $x \rightarrow x^*$, se

$$f(x) = h(x) + o(|h(x)|), \quad \text{quando } x \rightarrow x^*,$$

onde o símbolo (de Landau) $o(|h(x)|)$, $x \rightarrow x^*$, designa uma função genérica g tal que

$$\lim_{x \rightarrow x^*} \frac{|g(x)|}{|h(x)|} = 0.$$

Exemplo. $1 - \cos x \doteq \frac{x^2}{2}, \quad x \rightarrow 0.$

Proposição. Seja $\phi : I \subset \mathbb{R} \rightarrow \mathbb{R}$, $\phi \in C^2(I)$. Sejam $x \in I$ e $\tilde{x} \in I$ um valor que aproxima x . Então:

$$e_{\phi(\tilde{x})} = \phi(x) - \phi(\tilde{x}) \doteq \phi'(x) e_{\tilde{x}} =: e_{\phi(\tilde{x})}^L, \quad \text{quando } e_{\tilde{x}} \rightarrow 0,$$

e, no caso de ser $\phi(x) \neq 0$, $x \neq 0$,

$$\delta_{\phi(\tilde{x})} = \frac{e_{\phi(\tilde{x})}}{\phi(x)} \doteq p_{\phi}(x) \delta_{\tilde{x}} =: \delta_{\phi(\tilde{x})}^L, \quad \text{quando } \delta_{\tilde{x}} \rightarrow 0,$$

onde

$$p_{\phi}(x) = \frac{x \phi'(x)}{\phi(x)}.$$

Chama-se a $|p_\phi(x)|$ o **número de condição** de ϕ em x .

Exemplo. $\phi : \mathbb{R} \rightarrow \mathbb{R}$, $\phi(x) = x^m$, $m \in \mathbb{N}_1$

$$\delta_{\phi(\tilde{x})}^L = m\delta_{\tilde{x}}, \quad \delta_{\phi(\tilde{x})} = m\delta_{\tilde{x}} - \sum_{i=2}^m \binom{m}{i} (-\delta_{\tilde{x}})^i$$

Exemplo. Verificar que $\delta_{f \circ g(\tilde{x})}^L = p_f(g(x))p_g(x)\delta_{\tilde{x}}$.

Proposição. Seja $\phi : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $\phi \in C^2(D)$. Sejam $x \in D$ e $\tilde{x} \in D$ um valor que aproxima x . Então:

$$e_{\phi(\tilde{x})} = \phi(x) - \phi(\tilde{x}) \doteq e_{\phi(\tilde{x})}^L = \sum_{k=1}^n \frac{\partial \phi}{\partial x_k}(x) e_{\tilde{x}_k}, \quad \text{quando} \quad \sum_{k=1}^n |e_{\tilde{x}_k}| \rightarrow 0,$$

e, no caso de ser $\phi(x) \neq 0$, $\sum_{k=1}^n |x_k| \neq 0$,

$$\delta_{\phi(\tilde{x})} = \frac{e_{\phi(\tilde{x})}}{\phi(x)} \doteq \delta_{\phi(\tilde{x})}^L = \sum_{k=1}^n p_{\phi, x_k}(x) \delta_{\tilde{x}_k}, \quad \text{quando} \quad \sum_{k=1}^n |\delta_{\tilde{x}_k}| \rightarrow 0,$$

onde

$$p_{\phi, x_k}(x) = \frac{x_k \frac{\partial \phi}{\partial x_k}(x)}{\phi(x)}.$$

Exemplo. Operações aritméticas ($x, y \in \mathbb{R}$)

$\phi(x, y)$	$\delta_{\phi(\tilde{x}, \tilde{y})}^L$	$\delta_{\phi(\tilde{x}, \tilde{y})} - \delta_{\phi(\tilde{x}, \tilde{y})}^L$
$x + y$	$\frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}}$	0
$x - y$	$\frac{x}{x-y}\delta_{\tilde{x}} - \frac{y}{x-y}\delta_{\tilde{y}}$	0
$x \times y$	$\delta_{\tilde{x}} + \delta_{\tilde{y}}$	$-\delta_{\tilde{x}}\delta_{\tilde{y}}$
$x \div y$	$\delta_{\tilde{x}} - \delta_{\tilde{y}}$	$\frac{\delta_{\tilde{y}}(\delta_{\tilde{x}} - \delta_{\tilde{y}})}{1 - \delta_{\tilde{y}}}$

Exemplo. $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\phi(x, y) = x^2 - y^2$, $\delta_{\phi(\tilde{x}, \tilde{y})}^L = \frac{2}{x^2 - y^2}(x^2\delta_{\tilde{x}} - y^2\delta_{\tilde{y}})$

Propagação de erros em algoritmos

Definição. Uma função $\phi : I \subset \mathbb{R} \rightarrow \mathbb{R}$ diz-se uma **função elementar** num sistema \mathbb{F} se $\forall x \in I \cap \mathbb{F}$ o valor que aproxima $\phi(x)$ em \mathbb{F} é dado por

$$\tilde{\phi}(x) = \text{fl}(\phi(x)).$$

Proposição. Seja $\phi : I \subset \mathbb{R} \rightarrow \mathbb{R}$ uma função elementar num sistema \mathbb{F} . Seja \tilde{x} um valor aproximado de $x \in I$. Então:

(1)

$$e_{\tilde{\phi}(x)} = \phi(x) - \tilde{\phi}(x) = \phi(x) - \text{fl}(\phi(x)) = \phi(x)\delta_{\text{arr},\phi}$$

onde $|\delta_{\text{arr},\phi}| \leq u$, e u é a unidade de arredondamento de \mathbb{F} .

(2)

$$\delta_{\tilde{\phi}(\tilde{x})} \doteq \delta_{\tilde{\phi}(\tilde{x})}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr},\phi}, \quad \delta_{\phi(\tilde{x})}^L = p_{\phi}(x)\delta_{\tilde{x}}.$$

Dem.: (2)

$$\begin{aligned} e_{\tilde{\phi}(\tilde{x})} &= \phi(x) - \tilde{\phi}(\tilde{x}) \\ &= \phi(x) - \phi(\tilde{x}) + \phi(\tilde{x}) - \tilde{\phi}(\tilde{x}) \\ &\doteq \phi'(x)e_{\tilde{x}} + \phi(\tilde{x})\delta_{\text{arr},\phi} \\ &\doteq \phi'(x)e_{\tilde{x}} + \phi(x)\delta_{\text{arr},\phi} \end{aligned}$$

Nota. A definição de função elementar generaliza-se para funções de mais de uma variável e, em particular, para as operações aritméticas.

Definição. Por **algoritmo** entende-se uma sequência finita de operações elementares que conduz a um valor aproximado da solução de um problema.

Exemplo. Algoritmo para o cálculo de $z = \phi(x) = 1 - \cos x$, $x \in \mathbb{R}$:

$$u = \cos x = \theta(x); \quad z = 1 - u = \psi(u).$$

$$\text{Alg. 1: } u = \cos x = \theta(x), \quad z = 1 - u = \psi(u)$$

$$\text{Alg. 2: } u_1 = \frac{x}{2}, \quad u_2 = \sin u_1, \quad u_3 = u_2^2, \quad z = 2 \times u_3$$

Exemplo. Algoritmos para o cálculo de $z = \phi(x) = x_1^2 - x_2^2$, $x = (x_1, x_2) \in \mathbb{R}^2$:

$$\text{Alg. 1: } \begin{cases} u_1 = x_1 \times x_1 = \theta_1(x) \\ u_2 = x_2 \times x_2 = \theta_2(x) \\ z = u_1 - u_2 = \psi(u) \end{cases} \quad \text{Alg. 2: } \begin{cases} u_1 = x_1 + x_2 = \theta_1(x) \\ u_2 = x_1 - x_2 = \theta_2(x) \\ z = u_1 \times u_2 = \psi(u) \end{cases}$$

ψ, θ_1, θ_2 são em cada caso as funções elementares.

Exemplo. Algoritmo para o cálculo de $z = \phi(x)$, $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$u = \theta(x), \quad \theta : \mathbb{R}^n \rightarrow \mathbb{R}^p, \quad z = \psi(u), \quad \psi : \mathbb{R}^p \rightarrow \mathbb{R}.$$

Pondo $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ então $\theta_1, \dots, \theta_p, \psi$ são as $p + 1$ funções elementares.

Proposição. Suponhamos que o cálculo de $z = \phi(x)$, $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$, é efectuado por uma

algoritmo ϕ_* com $p + 1$ passos em que a cada passo corresponde uma função elementar a que está associado um certo erro de arredondamento. Seja \tilde{x} um valor aproximado de x . Então o erro relativo total de $\tilde{z} = \tilde{\phi}_*(\tilde{x})$ em relação a $\phi(x)$ é dado por

$$\delta_{\tilde{z}} \doteq \delta_{\tilde{z}}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr}}^L,$$

onde

$$\delta_{\phi(\tilde{x})}^L = \sum_{k=1}^n p_{\phi, x_k} \delta_{\tilde{x}_k}, \quad \delta_{\text{arr}}^L = \sum_{k=1}^{p+1} q_k \delta_{\text{arr}_k}.$$

Os pesos q_1, q_2, \dots, q_{p+1} dependem do algoritmo e $|\delta_{\text{arr}_k}| \leq u, \forall k \in \{1, 2, \dots, p+1\}$.

Exemplo. No caso dos dois algoritmos para o cálculo de $z = 1 - \cos x$ obtêm-se os seguintes erros relativos:

Algoritmo 1: $\delta_{\tilde{z}}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr}}^L$

$$\delta_{\phi(\tilde{x})}^L = p_{\phi}(x) \delta_{\tilde{x}} = \frac{x \sin x}{1 - \cos x} \delta_{\tilde{x}},$$

$$\delta_{\text{arr}}^L = \frac{-\cos x}{1 - \cos x} \delta_{\text{arr}, \theta} + \delta_{\text{arr}, \psi}$$

Algoritmo 2: $\delta_{\tilde{z}}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr}}^L$

$$\delta_{\phi(\tilde{x})}^L = p_{\phi}(x) \delta_{\tilde{x}},$$

$$\delta_{\text{arr}}^L = p_{\phi}(x) \delta_{\tilde{x}} \delta_{\text{arr}, 1} + 2\delta_{\text{arr}, 2} + \delta_{\text{arr}, 3} + \delta_{\text{arr}, 4}$$

Exemplo. No caso dos dois algoritmos para o cálculo de $z = x_1^2 - x_2^2$ obtêm-se os seguintes erros relativos:

Algoritmo 1: $\delta_{\tilde{z}}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr}}^L$

$$\delta_{\phi(\tilde{x})}^L = \frac{2}{z} (x_1^2 \delta_{\tilde{x}_1} - x_2^2 \delta_{\tilde{x}_2}),$$

$$\delta_{\text{arr}}^L = \frac{x_1^2}{z} \delta_{\text{arr}, \theta_1} - \frac{x_2^2}{z} \delta_{\text{arr}, \theta_2} + \delta_{\text{arr}, \psi}$$

Algoritmo 2: $\delta_{\tilde{z}}^L = \delta_{\phi(\tilde{x})}^L + \delta_{\text{arr}}^L$

$$\delta_{\phi(\tilde{x})}^L = \frac{2}{z} (x_1^2 \delta_{\tilde{x}_1} - x_2^2 \delta_{\tilde{x}_2}),$$

$$\delta_{\text{arr}}^L = \delta_{\text{arr}, \theta_1} + \delta_{\text{arr}, \theta_2} + \delta_{\text{arr}, \psi}$$

Condicionamento e estabilidade numérica

- Qualquer problema matemático pode ser descrito da seguinte forma:

Seja D o conjunto de dados do problema e seja S o conjunto de todas as soluções (resultados) possíveis do problema. Seja $f : D \rightarrow S$ a aplicação que a cada elemento de D associa um elemento de S , a solução do problema.

Definição. Um problema diz-se **estável** ou **bem posto** se a pequenos erros relativos dos dados correspondem pequenos erros relativos dos resultados. Caso contrário o problema diz-se **instável** ou **mal posto**. Em certos casos esta noção pode ser concretizada. Diz-se que o problema é **estável** para um dado $x \in D$ se existir uma constante $K \geq 0$ tal que é satisfeita a seguinte desigualdade:

$$\max_{1 \leq j \leq m} |\delta_{\tilde{z}_j}| \leq K \max_{1 \leq i \leq n} |\delta_{\tilde{x}_i}|, \quad \tilde{x} \in V_x,$$

onde $\tilde{z} = f(\tilde{x})$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, e $V_x \subset D$ é uma vizinhança de x . Nestes casos diz-se ainda que o problema é **bem condicionado** se K é pequeno e **mal condicionado** se K é grande.

Exemplo. Vimos que para $z = f(x)$, $\tilde{z} = f(\tilde{x})$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\delta_{\tilde{z}} \doteq \delta_{\tilde{z}}^L = \sum_{i=1}^n p_{f,x_i} \delta_{\tilde{x}_i}.$$

O problema é pois mal posto para $x \in D$ se algum dos números de condição p_{f,x_i} tender para infinito para esse x . Caso isto não aconteça podemos definir

$$K = \sum_{i=1}^n |p_{f,x_i}|.$$

- A resolução numérica deste problema significa que existe uma outra aplicação $f_* : D \rightarrow S$, que a cada elemento de D associa um elemento de S , a solução numérica do problema. Neste caso para além dos erros dos dados temos de considerar os erros de arredondamento.

Definição. Um algoritmo diz-se **numericamente (ou computacionalmente) estável** se a pequenos erros relativos dos dados e a pequenos valores da unidade de arredondamento correspondem pequenos erros relativos nos resultados do algoritmo. Caso contrário o algoritmo diz-se **numericamente (ou computacionalmente) instável**. Em certos casos esta noção pode ser concretizada. Diz-se que o algoritmo é **numericamente estável** para um dado $x \in D$ se existir uma constante $K \geq 0$ tal que é satisfeita a seguinte desigualdade:

$$\max_{1 \leq j \leq m} |\delta_{\tilde{z}_j}| \leq K \left(\max_{1 \leq i \leq n} |\delta_{\tilde{x}_i}| + u \right), \quad \tilde{x} \in V_x,$$

onde $\tilde{z} = \tilde{f}_*(\tilde{x})$ e $V_x \subset D$ é uma vizinhança de x . (Recorde-se que a unidade de arredondamento é um majorante de todos os erros relativos de arredondamento.)

Exemplo. Vimos que para $z = f(x)$, $\tilde{z} = \tilde{f}_*(\tilde{x})$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\delta_{\tilde{z}} \doteq \delta_{\tilde{z}}^L = \sum_{i=1}^n p_{f,x_i} \delta_{\tilde{x}_i} + \sum_{i=1}^{p+1} q_i \delta_{arr_i}.$$

O algoritmo é pois numericamente instável para $x \in D$ se algum dos números de condição p_{f,x_i} ou algum dos q_i tender para infinito para esse x . Caso isto não aconteça podemos definir

$$K = \max \left\{ \sum_{i=1}^n |p_{f,x_i}|, \sum_{i=1}^{p+1} |q_i| \right\}.$$

Exemplo. O problema de cálculo de $z = 1 - \cos x$ é um problema mal posto para $x = 2k\pi$, $k \in \mathbf{Z} \setminus \{0\}$. O Algoritmo 2 é numericamente instável para os mesmos valores de x . O Algoritmo 1 é também numericamente instável para $x = 0$.