MATEMÁTICA COMPUTACIONAL

Apontamentos das aulas

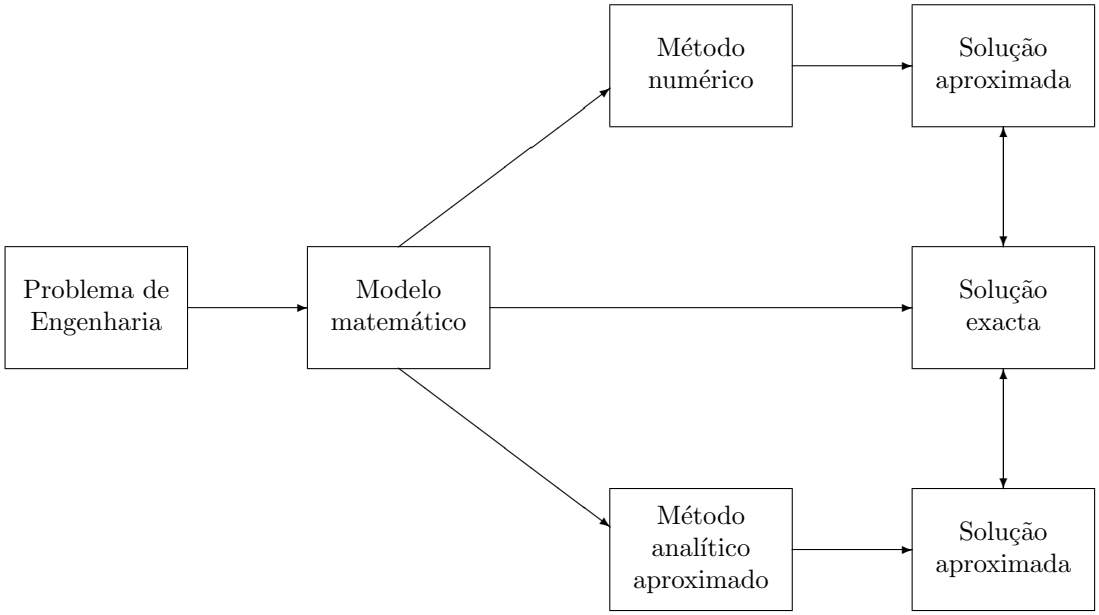
Filipe J. Romeiras

Departamento de Matemática
Instituto Superior Técnico

Junho de 2008

CONTEÚDO

1. REPRESENTAÇÃO DE NÚMEROS E TEORIA DE ERROS



Tipos de erro

• **Erros inerentes**:
◇ modelo matemático incompleto;
◇ erros nos dados, parâmetros e constantes matemáticas;
◇ exemplos:
(i) corpo em movimento vertical na atmosfera terrestre sujeito à força de atracção gravitacional da Terra e à força de resistência do ar;
(ii) a reacção de Belousov-Zhabotinski (oscilações químicas);
(iii) circuitos electrónicos não-lineares.
• **Erros de método (ou de truncatura ou de discretização)**:
◇ exemplos:
(i) cálculo do valor aproximado da raiz de uma função pelo método de Newton;
(ii) 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 $x \neq 0$: $\delta_{\tilde{x}} = \dfrac{e_{\tilde{x}}}{x}$, 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.

$$x = \frac{1}{3}, \qquad \tilde{x} = 0.3333, \qquad y = \frac{1}{3000}, \qquad \tilde{y} = 0.0003.$$
$$e_{\tilde{x}} = e_{\tilde{y}} = 0.0000333\ldots, \qquad \delta_{\tilde{x}} \approx 0.0001(0.01\%) \qquad \delta_{\tilde{y}} \approx 0.1(10\%)$$

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

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

onde

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

Simbolicamente escreve-se

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

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

Com efeito:

$$\begin{aligned}
198 &= 2 \times 99 = 2 \times (2 \times 49 + 1) = 2 \times (2 \times (2 \times 24 + 1) + 1) = \\
&= 2 \times (2 \times (2 \times (2 \times 12) + 1) + 1) = 2 \times (2 \times (2 \times (2 \times (2 \times 6)) + 1) + 1) = \\
&= 2 \times (2 \times (2 \times (2 \times (2 \times (2 \times 3))) + 1) + 1) = \\
&= 2 \times (2 \times (2 \times (2 \times (2 \times (2 \times (2 + 1)))) + 1) + 1) = \\
&= 2^7 + 2^6 + 2^2 + 2^1 = (11000110)_2
\end{aligned}$$

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. $N = 31 : \quad 2^N - 1 = 2.147.483.647$

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

$$x = \sigma(d_m\beta^m + d_{m-1}\beta^{m-1} + \cdots + d_1\beta^1 + d_0\beta^0 +$$
$$+d_{-1}\beta^{-1} + d_{-2}\beta^{-2} + \cdots d_{-n}\beta^{-n} + \cdots)$$

onde

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

Simbolicamente escreve-se

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

ou

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

Exemplo.

$$19 = (10011)_2, \qquad 0.875 = (0.111)_2,$$
$$19.875 = (10011.111)_2 = (0.10011111)_2 \times 2^{(101)_2}$$
$$0.1 = (0.11001100\ldots)_2 \times 2^{-(11)_2}$$

Com efeito:

$$19 = 2 \times 9 + 1 = 2 \times (2 \times 4 + 1) + 1 = 2 \times (2 \times (2 \times 2) + 1) + 1 = 2^4 + 2^1 + 2^0$$

$$0.875 = d_{-1} \times 2^{-1} + d_{-2} \times 2^{-2} + d_{-3} \times 2^{-3} + d_{-4} \times 2^{-4} + \ldots$$
$$1.750 = d_{-1} + d_{-2} \times 2^{-1} + d_{-3} \times 2^{-2} + d_{-4} \times 2^{-3} + \ldots \quad \Rightarrow \quad d_{-1} = 1$$
$$0.750 = d_{-2} \times 2^{-1} + d_{-3} \times 2^{-2} + d_{-4} \times 2^{-3} + \ldots$$
$$1.5 = d_{-2} + d_{-3} \times 2^{-1} + d_{-4} \times 2^{-2} + \ldots \quad \Rightarrow \quad d_{-2} = 1$$
$$0.5 = d_{-3} \times 2^{-1} + d_{-4} \times 2^{-2} + \ldots$$
$$1.0 = d_{-3} + d_{-4} \times 2^{-1} + \ldots \quad \Rightarrow \quad d_{-3} = 1, \qquad d_{-4} = d_{-5} = \cdots = 0$$

• Notação científica

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

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

$$\text{(mantissa) } m = (0.a_1a_2\ldots)_\beta \in [\beta^{-1}, 1[, \qquad a_i \in \{0,1,\ldots,\beta-1\}, \qquad a_1 \neq 0$$

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

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

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

$$m = (0.a_1a_2\ldots a_n)_\beta \in [\beta^{-1}, 1-\beta^{-n}], \qquad a_i \in \{0,1,\ldots,\beta-1\}, \qquad a_1 \neq 0$$

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

Proposição.

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

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

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

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

Exemplo: $\mathrm{FP}(2, 3, -1, 1)$, cujos elementos positivos são:

$$\begin{array}{lll}
(0.100)_2 \times 2^{-1} = \dfrac{4}{16} & (0.100)_2 \times 2^{0} = \dfrac{8}{16} & (0.100)_2 \times 2^{1} = \dfrac{16}{16} \\
(0.101)_2 \times 2^{-1} = \dfrac{5}{16} & (0.101)_2 \times 2^{0} = \dfrac{10}{16} & (0.101)_2 \times 2^{1} = \dfrac{20}{16} \\
(0.110)_2 \times 2^{-1} = \dfrac{6}{16} & (0.110)_2 \times 2^{0} = \dfrac{12}{16} & (0.110)_2 \times 2^{1} = \dfrac{24}{16} \\
(0.111)_2 \times 2^{-1} = \dfrac{7}{16} & (0.111)_2 \times 2^{0} = \dfrac{14}{16} & (0.111)_2 \times 2^{1} = \dfrac{28}{16}
\end{array}$$

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

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

$$L^- = 2^{-126} \approx 0.118 \times 10^{-37}, \qquad L^+ = (1 - 2^{-24}) \times 2^{128} \approx 0.340 \times 10^{39},$$

$$N = 254 \times 2^{23} \approx 0.213 \times 10^{10}$$

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

$$L^- = 2^{-1022} \approx 0.223 \times 10^{-307}, \qquad L^+ = (1 - 2^{-53}) \times 2^{1024} \approx 0.180 \times 10^{309},$$

$$N = 2046 \times 2^{52} \approx 0.921 \times 10^{19}$$

Arredondamentos

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

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

$$x = \sigma(0.a_1 a_2 \ldots a_n a_{n+1} \ldots)_\beta \times \beta^t,$$

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

(i) Arredondamento por corte:

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

(ii) Arredondamento simétrico (β par):

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

ou, de uma forma equivalente,

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

Nota.

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

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

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

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

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

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

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

Erros de arredondamento

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

(i) Arredondamento por corte:

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

(ii) Arredondamento simétrico:

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

Dem.:

$$x = \sigma(0.a_1a_2\ldots a_na_{n+1}\ldots)_\beta \times \beta^t, \qquad |x| \geq \beta^{-1+t}$$

(i) Arredondamento por corte:

$$\tilde{x} = \mathrm{fl_c}(x) = \sigma(0.a_1a_2\ldots a_n)_\beta \times \beta^t$$
$$e_{\tilde{x}} = x - \tilde{x} = \sigma(0.0\ldots 0a_{n+1}\ldots)_\beta \times \beta^t = \sigma(0.a_{n+1}a_{n+2}\ldots)_\beta \times \beta^{t-n}$$
$$|e_{\tilde{x}}| < \beta^{t-n}$$
$$|\delta_{\tilde{x}}| = \frac{|e_{\tilde{x}}|}{|x|} < \beta^{1-n}$$

(ii) Arredondamento simétrico:

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

$$\tilde{x} = \mathrm{fl_s}(x) = \sigma(0.a_1a_2\ldots a_n)_\beta \times \beta^t$$
$$e_{\tilde{x}} = \sigma(0.a_{n+1}a_{n+2}\ldots)_\beta \times \beta^{t-n}$$
$$|e_{\tilde{x}}| < \frac{\beta}{2}\,\beta^{-1}\,\beta^{t-n} = \frac{1}{2}\,\beta^{t-n}$$
$$|\delta_{\tilde{x}}| < \frac{1}{2}\,\beta^{1-n}$$

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

$$\tilde{x} = \mathrm{fl_s}(x) = \sigma\left[(0.a_1a_2\ldots a_n)_\beta + \beta^{-n}\right] \times \beta^t$$
$$e_{\tilde{x}} = \sigma\left[(0.a_{n+1}a_{n+2}\ldots)_\beta - 1\right] \times \beta^{t-n}$$
$$|e_{\tilde{x}}| \leq \frac{1}{2}\,\beta^{t-n}$$
$$|\delta_{\tilde{x}}| < \frac{1}{2}\,\beta^{1-n}$$

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

Definição. Dado um sistema FP(β, n, t^-, t^+) define-se a **unidade de arredondamento do sistema** por

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

Exemplo. Sistemas definidos pela Norma IEC559.

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

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

Operações aritméticas num sistema de ponto flutuante

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

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

Exemplo. Sendo $x = \pi$ e $y = \dfrac{333}{106}$ e considerando um sistema de ponto flutuante FP(10,

6, -10, 10) com arredondamento simétrico obtém-se:

$$x \boxed{+} y = 0.628310 \times 10, \qquad x \boxed{-} y = 0.800000 \times 10^{-4},$$

$$x \boxed{\times} y = 0.986934 \times 10, \qquad x \boxed{\div} y = 0.100003 \times 10$$

Note-se que:

$$x = \pi = 3.14159265358\ldots, \qquad \tilde{x} = 0.314159 \times 10$$
$$y = \frac{333}{106} = 3.14150943396\ldots, \qquad \tilde{y} = 0.314151 \times 10$$

Algarismos significativos

Definição. Seja

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

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 $\pi = 3.14159265358\ldots$, então:

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

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

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

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

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), $\qquad u \approx 0.596046 \times 10^{-7}$

6 ou 7 algarismos significativos

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

$$x = \sigma m 10^t, \qquad x - \tilde{x} = x\delta_{\tilde{x}}, \qquad |\delta_{\tilde{x}}| < u = m_u 10^{t_u}$$
$$|x - \tilde{x}| < m m_u 10^{t+t_u}, \qquad 0.1 m_u \le m m_u < m_u$$

Erros de "overflow" e "underflow"

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

Exemplo. Cálculo de $|x + iy|$ para

(i) $x = y = 10^{20}$; (ii) $x = 10^{20},\ y = 1$;

no sistema FP(2, 24, -125, 128).

$$|x+iy| = \sqrt{x^2+y^2} = \begin{cases} |x|\sqrt{1+\left(\dfrac{y}{x}\right)^2}, & |x| \ge |y| \\ |y|\sqrt{1+\left(\dfrac{x}{y}\right)^2}, & |x| < |y| \end{cases}$$

(i) $x = 10^{20} = y$: $\quad |x+iy| = 10^{20}\sqrt{2}$

(ii) $x = 10^{20},\ y = 1$: $\quad |x+iy| = 10^{20}\sqrt{1+\left(\dfrac{1}{10^{20}}\right)^2} \approx 10^{20}$

Exemplo. Cálculo das raízes de $ax^2 + bx + c = 0$ para

(i) $a = 10^{20},\ \ b = -3 \times 10^{20},\ \ c = 2 \times 10^{20}$;

(ii) $a = 2,\ \ b = -3 \times 10^{20},\ \ c = 2$;

no sistema FP(2, 24, -125, 128).

$$x_- = -\frac{1}{2a}\left(b + \sqrt{b^2 - 4ac}\right), \qquad x_+ = -\frac{1}{2a}\left(b - \sqrt{b^2 - 4ac}\right)$$

(i) As raízes coincidem com as do caso $\ \bar{a} = 1,\ \ \bar{b} = -3,\ \ \bar{z} = 2$:

$$x_- = 1, \qquad x_+ = 2$$

(ii) $\quad x_- = -\dfrac{b}{2a}\left(1 + \sqrt{1 - \dfrac{4ac}{b^2}}\right), \qquad x_+ = \dfrac{c/a}{x_-}$

$$x_- = \frac{3 \times 10^{20}}{4}\left(1 + \sqrt{1 - \frac{16}{(3 \times 10^{20})^2}}\right) \approx \frac{3}{2} \times 10^{20}, \qquad x_+ \approx \frac{2}{3} \times 10^{-20}$$

Propagação de erros (teoria linearizada)

Definição. Diz-se que

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

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

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

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

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

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

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

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

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

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

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

$$\delta^L_{\phi(\tilde{x})} = m\delta_{\tilde{x}}, \qquad \delta_{\phi(\tilde{x})} = m\delta_{\tilde{x}} - \sum_{i=2}^{m} \binom{m}{i} (-\delta_{\tilde{x}})^i$$

Exemplo. $\delta^L_{f\circ g(\tilde{x})} = p_f(g(x))p_g(x)\delta_{\tilde{x}}$.

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

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

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

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

onde

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

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

$\phi(x,y)$	$\delta^L_{\phi(\tilde{x},\tilde{y})}$	$\delta_{\phi(\tilde{x},\tilde{y})} - \delta^L_{\phi(\tilde{x},\tilde{y})}$
$x + y$	$\frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}}$	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. $\phi : \mathbb{R}^2 \to \mathbb{R}, \;\; \phi(x,y) = x^2 - y^2, \qquad \delta^L_{\phi(\tilde{x},\tilde{y})} = \frac{2}{x^2 - y^2}(x^2\delta_{\tilde{x}} - y^2\delta_{\tilde{y}})$

Propagação de erros em algoritmos

Definição. Uma função $\phi : I \subset \mathbb{R} \to \mathbb{R}$ 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

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

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

(1)

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

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

(2)

$$\delta_{\tilde{\phi}(\tilde{x})} \doteq \delta^L_{\tilde{\phi}(\tilde{x})} = \delta^L_{\phi(\tilde{x})} + \delta_{\mathrm{arr},\phi}, \qquad \delta^L_{\phi(\tilde{x})} = p_\phi(x)\delta_{\tilde{x}}.$$

Dem.: (2)

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

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

Alg. 1: $u = \cos x = \theta(x), \qquad z = 1 - u = \psi(u)$

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

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

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

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

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

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

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

Proposição. Suponhamos que o cálculo de $z = \phi(x), \ \ \phi : \mathbb{R}^n \to \mathbb{R}$, 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

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

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

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

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

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

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

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

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

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

Com efeito:

Algoritmo 1:

$$\delta_{\tilde{u}}^L = p_\theta(x)\delta_{\tilde{x}} + \delta_{\text{arr},\theta}, \qquad p_\theta(x) = \frac{x\theta'(x)}{\theta(x)} = \frac{-x\sin x}{\cos x}$$

$$\delta_{\tilde{z}}^L = p_\psi(u)\delta_{\tilde{u}}^L + \delta_{\text{arr},\psi}, \qquad p_\psi(u) = \frac{u\psi'(u)}{\psi(u)} = \frac{-u}{1-u} = \frac{-\cos x}{1-\cos x}$$

$$\delta_{\tilde{z}}^L = p_\psi(u)\left[p_\theta(x)\delta_{\tilde{x}} + \delta_{\text{arr},\theta}\right] + \delta_{\text{arr},\psi} = p_\psi(u)p_\theta(x)\delta_{\tilde{x}} + p_\psi(u)\delta_{\text{arr},\theta} + \delta_{\text{arr},\psi}$$

Algoritmo 2:

$$\delta_{\tilde{u}_1}^L = \delta_{\tilde{x}} + \delta_{\text{arr},1}$$

$$\delta_{\tilde{u}_2}^L = p_{u_2}(u_1)\delta_{\tilde{u}_1}^L + \delta_{\text{arr},2}, \qquad p_{u_2}(u_1) = \frac{u_1\cos u_1}{u_2}$$

$$\delta_{\tilde{u}_3}^L = 2\,\delta_{\tilde{u}_2}^L + \delta_{\text{arr},3}$$

$$\delta_{\tilde{z}}^L = \delta_{\tilde{u}_3}^L + \delta_{\text{arr},4} = 2\left[p_{u_2}(u_1)\left(\delta_{\tilde{x}} + \delta_{\text{arr},1}\right) + \delta_{\text{arr},2}\right] + \delta_{\text{arr},3} + \delta_{\text{arr},4}$$

$$2p_{u_2}(u_1) = p_\phi(x)$$

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

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

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

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

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

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

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 $f : D \to S$ 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:

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

onde $\tilde{z} = f(\tilde{x}), \quad f : \mathbb{R}^n \to \mathbb{R}^m$, 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 $z = f(x), \;\; \tilde{z} = f(\tilde{x}), \;\; f : \mathbb{R}^n \to \mathbb{R}$,

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

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 $f_* : D \to S$, 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:

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

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 $z = f(x), \;\; \tilde{z} = \tilde{f}_*(\tilde{x}), \;\; f : \mathbb{R}^n \to \mathbb{R}$,

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

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 $x = 2k\pi, \quad k \in \mathbb{Z} \setminus \{0\}$. O Algoritmo 2 é numericamente instável para os mesmos valores de x. O Algoritmo 1 é também numericamente instável para $x = 0$.

Com efeito:

Algoritmo 1: $\delta_{\tilde{z}}^L = p_\phi(x)\delta_{\tilde{x}} + q(x)\delta_{\mathrm{arr},\theta} + \delta_{\mathrm{arr},\psi}$

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

$$p_\phi(x) = \frac{x \sin x}{1 - \cos x}, \qquad q(x) = \frac{-\cos x}{1 - \cos x}$$

p_ϕ é singular para $x = 2k\pi, \;\; k \in \mathbb{Z}$; $\quad \lim_{x \to 0} p_\phi(x) = 2.$

q é singular para $x = 2k\pi, \;\; k \in \mathbb{Z}$.

2. MÉTODOS ITERATIVOS

Normas vectoriais

Definição. Seja E um espaço vectorial sobre $\mathbb{R}$ (ou $\mathbb{C}$). Uma função $N : E \to \mathbb{R}$ diz-se uma **norma** se verificar as seguintes condições:

(1) $N(x) \geq 0, \ \ \forall x \in E; \ \ N(x) = 0 \Leftrightarrow x = 0.$

(2) $N(\alpha x) = |\alpha| N(x), \ \ \forall x \in E, \ \ \forall \alpha \in \mathbb{R} \ \ (\text{ou } \ \mathbb{C}).$

(3) $N(x + y) \leq N(x) + N(y), \ \ \forall x, y \in E.$ (Desigualdade triangular)

Notação. $\|x\|_N = N(x)$

Exemplo. Normas em $\mathbb{R}^n$ (ou $\mathbb{C}^n$). Seja $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

(1) Norma da soma
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

(2) Norma Euclidiana
$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

(3) Norma do máximo
$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

(4) Norma-p, $\ p \geq 1$
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Nota. A desigualdade triangular para a norma-p,
$$\|x + y\|_p \leq \|x\|_p + \|y\|_p, \qquad \forall x, y \in \mathbb{C}^n, \quad \forall p \geq 1,$$
é conhecida por **desigualdade de Minkowski**.

Definição. Um espaço vectorial no qual está definida uma norma diz-se um **espaço normado**.

Definição. Sejam E um espaço normado com a norma N e X um subconjunto de E. Então $f : E \to \mathbb{R}$ diz-se uma **função contínua** em X se e só se
$$\forall \varepsilon > 0 \ \ \exists \delta > 0 \ \ \forall x, y \in X : N(x - y) < \delta \Rightarrow |f(x) - f(y)| < \varepsilon.$$

Proposição. Uma norma N num espaço vectorial E é uma função contínua de E em $\mathbb{R}$.

Definição. Diz-se que duas normas N e M num espaço vectorial E são **equivalentes** se existirem constantes positivas C_1 e C_2 tais que

$$C_1 M(x) \leq N(x) \leq C_2 M(x), \qquad \forall x \in E.$$

Teorema. Todas as normas num espaço vectorial de dimensão finita são equivalentes.

Exemplo. Sendo $x \in \mathbb{C}^n$:

(1) $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$

(2) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$

(3) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$

Com efeito:

(1) $\|x\|_\infty = |x_m| \leq \sum_i |x_i| = \|x\|_1 \leq n|x_m| = n\|x\|_\infty$

(2) $\|x\|_\infty^2 = |x_m|^2 \leq \sum_i |x_i|^2 = \|x\|_2^2 \leq n|x_m|^2 = n\|x\|_\infty^2$

(3) $\|x\|_2^2 = \sum_i |x_i|^2 \leq \left(\sum_i |x_i|\right)^2 = \|x\|_1^2$

$$\|x\|_1 = \sum_i 1 \cdot |x_i| \leq \left(\sum_i 1^2\right)^{1/2} \left(\sum_i |x_i|^2\right)^{1/2} = \sqrt{n}\|x\|_2$$

(onde se usou a desigualdade de Cauchy-Schwarz)

Definição. Sendo $x, y \in \mathbb{C}^n$ define-se o seu **produto interno** $< x, y >$ por

$$< x, y >= y^* x = \sum_{i=1}^n x_i \bar{y}_i.$$

Proposição. Sendo $x, y, z \in \mathbb{C}^n$, $\alpha \in \mathbb{C}$, $A \in \mathbb{M}^n(\mathbb{C})$.

(1) $< x, y + z >=< x, y > + < x, z >$;

(2) $< \alpha x, y >= \alpha < x, y >$; $\qquad < x, \alpha y >= \bar{\alpha} < x, y >$;

(3) $< x, y >= \overline{< y, x >}$;

(4) $< x, x >\geq 0$; $\qquad < x, x >= 0 \Leftrightarrow x = 0$;

(5) $\|x\|_2 = (< x, x >)^{1/2}$;

(6) $| < x, y > | \leq \|x\|_2 \|y\|_2$, verificando-se a igualdade se e só se $y = \alpha x$, $\alpha \in \mathbb{C}$.
(**Desigualdade de Cauchy-Schwarz**)

(7) $< Ax, y > = < x, A^* y >$, onde A^* designa a matrix conjugada transposta de A.

Nota. Designaremos o espaço de matrizes quadradas de ordem n por $\mathbb{M}^n(\mathbb{R})$ ou $\mathbb{M}^n(\mathbb{C})$, consoante os elementos das matrizes sejam reais ou complexos.

Erro, erro absoluto, erro relativo ($\tilde{x} \approx x \in E$)

Definição. Seja E um espaço normado com norma $\|\cdot\|$ e $\tilde{x}$ o valor aproximado de $x \in E$. 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 $x \neq 0$: $\delta_{\tilde{x}} = \dfrac{e_{\tilde{x}}}{\|x\|}$, ou $\|\delta_{\tilde{x}}\| = \dfrac{\|e_{\tilde{x}}\|}{\|x\|}$

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

Nota. Os erros dependem da norma utilizada.

Exemplo. Sendo $x = (\pi, \sqrt{2}, 1)$, $\tilde{x} = (3.14, 1.4, 1.01)$:

$$x - \tilde{x} = (0.00159265,\ \ 0.0142136,\ \ -0.01)$$

$$\|e_{\tilde{x}}\|_1 = 0.0258063, \qquad \|e_{\tilde{x}}\|_2 = 0.0174517, \qquad \|e_{\tilde{x}}\|_\infty = 0.0142136.$$

Convergência e ordem de convergência

Definição. Seja $\{x^{(k)}\}_{k\in\mathbb{N}}$ uma sucessão de elementos de um espaço normado E. Diz-se que esta sucessão **converge** para um certo $x \in E$ segundo a norma N, escrevendo-se simbolicamente $x^{(k)} \xrightarrow{N} x$, se e só se

$$\lim_{k\to\infty} \|x^{(k)} - x\|_N = 0.$$

Nota. Do teorema anterior sobre a equivalência de todas as normas num espaço de dimensão finita resulta que se uma dada sucessão $\{x^{(k)}\}$ converge para um certo $x \in E$ segundo uma certa norma N, então também converge para x segundo qualquer outra norma M. Esta observação permite-nos, enquanto estivermos a tratar de espaços de dimensão finita, falar de convergência sem nos referirmos a nenhuma norma em especial. Por isso dizemos simplesmente que a sucessão $\{x^{(k)}\}$ converge para x e escrevemos simbolicamente $x^{(k)} \to x$.

Nota. Sempre que não haja risco de confusão designaremos, por simplicidade de notação, o termo geral de uma sucessão por x_n em vez de por $x^{(n)}$.

Definição. Seja x_n uma sucessão convergente para x e seja $e_n = x - x_n \neq 0$. Para $r \geq 1$ consideremos a sucessão

$$K_n^{[r]} = \frac{\|e_{n+1}\|}{\|e_n\|^r}.$$

(1) Se existir n_0 tal que

$$n > n_0 \;\Rightarrow\; K_n^{[r]} \leq K^+ < M_r,$$

onde $M_1 = 1$ e $M_r = +\infty, \quad r > 1$, diz-se que a sucessão tem **ordem de convergência (o.c.) maior ou igual a** r.

(2) Se além disso

$$n > n_0 \;\Rightarrow\; 0 < K^- \leq K_n^{[r]},$$

diz-se que a sucessão tem **o.c.** r.

(3) Se existir

$$K_\infty^{[r]} = \lim_{n\to\infty} K_n^{[r]},$$

este valor é designado por **coeficiente assimptótico de convergência de ordem** r, verificando-se então:

$K_\infty^{[r]} < M_r$:	a sucessão tem o.c. maior ou igual a r;
$0 < K_\infty^{[r]} < M_r$:	a sucessão tem o.c. r; se $r = 1$ a convergência diz-se linear; se $r = 2$ a convergência diz-se quadrática;
$K_\infty^{[r]} = 0, \;\; \forall r > 1$:	a sucessão tem convergência exponencial;
$K_\infty^{[1]} = 0$:	a sucessão tem convergência supralinear;
$K_\infty^{[1]} = 1$:	a sucessão tem convergência logarítmica ou infralinear.

Nota. Se a sucessão tem o.c. r, então tem o.c. maior ou igual a $q, \quad 0 < q < r$. Com efeito, como $\lim_{n\to\infty} \|e^{(n)}\|^{r-q} = 0$, tem-se

$$\frac{\|e_{n+1}\|}{\|e_n\|^q} = \frac{\|e_{n+1}\|}{\|e_n\|^r}\|e_n\|^{r-q} \to K_\infty^{[r]} \times 0 = 0.$$

Por outro lado, se a sucessão tem o.c. r, então não pode ter o.c. superior a r. Com efeito, sendo $q > p$, como $\lim_{n\to\infty} \|e_n\|^{r-q} = \infty$ e $K_\infty^{[r]} \neq 0$, tem-se

$$\frac{\|e_{n+1}\|}{\|e_n\|^q} = \frac{\|e_{n+1}\|}{\|e_n\|^r}\|e_n\|^{r-q} \to K_\infty^{[r]} \times \infty = \infty.$$

Exemplo. As quatro sucessões seguintes convergem para 1 com a ordem de convergência

indicada ($a, b \in \mathbb{Z},\ \ a, b \geq 2$):

(1) $u_n = 1 + \dfrac{1}{n^a}$: convergência logarítmica ou infralinear

(2) $v_n = 1 + \dfrac{1}{a^n}$: convergência linear

(3) $x_n = 1 + \dfrac{1}{a^{b^n}}$: convergência (supralinear) de ordem b

(4) $y_n = 1 + \dfrac{1}{a^{b^{n^2}}}$: convergência exponencial

Com efeito:

$$\frac{|1-u_{n+1}|}{|1-u_n|} = \left(\frac{n}{n+1}\right)^a \underset{n\to\infty}{\longrightarrow} 1$$

$$\frac{|1-v_{n+1}|}{|1-v_n|} = \frac{1}{a} < 1$$

$$\frac{|1-x_{n+1}|}{|1-x_n|^r} = a^{b^n(r-b)} \underset{n\to\infty}{\longrightarrow} \begin{cases} 0, & r < b \\ 1, & r = b \\ \infty, & b < r \end{cases}$$

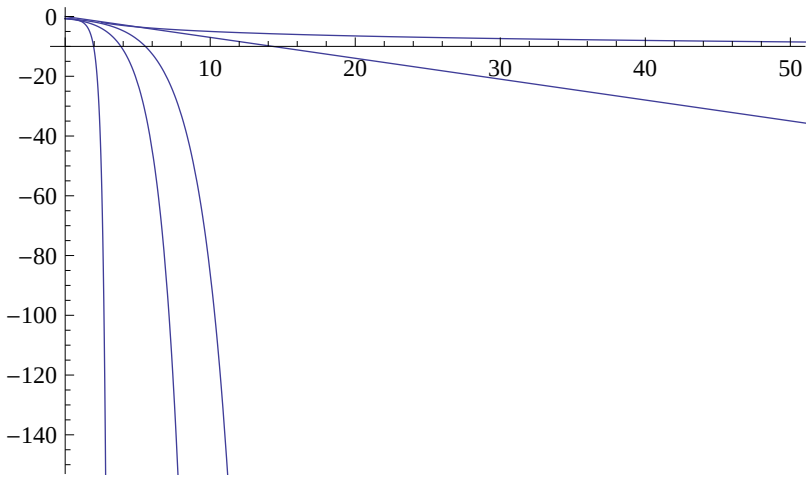
$$\frac{|1-y_{n+1}|}{|1-y_n|^r} = a^{b^{n^2}(r-b^{2n+1})} \underset{n\to\infty}{\longrightarrow} 0$$

Gráficos de $\log_{10}|e_n|$, $\ e_n = 1 - u_n$, em função de n para

$$(1)\ u_n = 1 + n^{-5}; \qquad (2)\ u_n = 1 + 5^{-n};$$

$$(3)'\ u_n = 1 + 5^{-r^n},\ \ r = \frac{1+\sqrt{5}}{2}; \qquad (3)\ u_n = 1 + 5^{-2^n}; \qquad (4)\ u_n = 1 + 5^{-2^{n^2}},$$

no sentido dos ponteiros do relógio.



Exemplo. Consideremos a sucessão

$$w_n = 1 + \frac{2 + (-1)^n}{4^n}.$$

Uma vez que

$$K_n^{[1]} = \frac{|1 - w_{n+1}|}{|1 - w_n|} = \frac{1}{4}\,\frac{2 + (-1)^{n+1}}{2 + (-1)^n} = \begin{cases} \dfrac{1}{12}, & n \text{ par}, \\ \dfrac{3}{4}, & n \text{ ímpar}, \end{cases}$$

não existe $K_\infty^{[1]}$. No entanto, como podemos escrever,

$$0 < \frac{1}{12} \leq K_n^{[1]} \leq \frac{3}{4} < 1,$$

concluímos que a sucessão tem convergência linear.

Métodos iterativos

Definição. Seja X um subconjunto de um espaço normado E e seja $p \geq 1$ um inteiro. Chama-se **método iterativo a p passos** em E à aplicação $\Psi : X^p \to E^{\mathbb{N}}$ que a cada $(\xi_0, \xi_1, \ldots, \xi_{p-1}) \in X^p$ associa uma sucessão $\{x^{(m)}\}_{m\in\mathbb{N}}$ de elementos de E tal que

$$\begin{cases} x^{(i)} = \xi_i, & \forall i \in \{0, 1, \ldots, p-1\} \\ x^{(m+p)} = \phi\left(x^{(m)}, x^{(m+1)}, \ldots, x^{(m+p-1)}\right), & \forall m \in \mathbb{N}, \end{cases}$$

onde $\phi : X^p \to X$ é uma função conhecida. Diz-se que ϕ é a **função iteradora**, $x^{(m+p)}$ é a **iterada** de ordem $m + p$ e $(\xi_0, \xi_1, \ldots, \xi_{p-1})$ são os **valores iniciais**.

Definição. Diz-se que o método iterativo Ψ a p passos é **convergente** para $x \in E$, num certo domínio $D \subset X^p$, se para quaisquer valores iniciais $(\xi_0, \xi_1, \ldots, \xi_{p-1}) \in D$ se verificar que $x^{(m)} \to x$.

Definição. Diz-se que o método iterativo é de ordem r se a sucessão gerada pelo método tem ordem de convergência r.

Exemplo. Os seguintes métodos iterativos são utilizados na determinação de zeros de funções $f : I \subset \mathbb{R} \to \mathbb{R}$:

(1) Método do ponto fixo ($f(x) = 0 \;\Leftrightarrow\; x = g(x)$):

$$\begin{cases} x_{m+1} = g(x_m), & m \in \mathbb{N}, \\ x_0 = \xi_0 \in I \end{cases}$$

$$\Psi : I \to \mathbb{R}^{\mathbb{N}}, \qquad \phi : I \to I, \qquad \phi(x) = g(x)$$

(2) Método de Newton:

$$\begin{cases} x_{m+1} = x_m - \dfrac{f(x_m)}{f'(x_m)}, & m \in \mathbb{N}, \\ x_0 = \xi_0 \in I \end{cases}$$

$$\Psi : I \to \mathbb{R}^{\mathbb{N}}, \qquad \phi : I \to I, \qquad \phi(x) = x - \frac{f(x)}{f'(x)}$$

(3) Método da secante:

$$\begin{cases} x_{m+2} = x_{m+1} - f(x_{m+1})\,\dfrac{x_{m+1} - x_m}{f(x_{m+1}) - f(x_m)}, & m \in \mathbb{N}, \\ x_0 = \xi_0 \in I, \;\; x_1 = \xi_1 \in I \end{cases}$$

$$\Psi : I^2 \to \mathbb{R}^{\mathbb{N}}, \qquad \phi : I^2 \to I, \qquad \phi(x,y) = \frac{xf(y) - yf(x)}{f(y) - f(x)}$$

Os métodos têm o.c. 1, 2 e $r = \dfrac{1+\sqrt{5}}{2}$, respectivamente.

O seguinte método iterativo utiliza-se na resolução de sistemas lineares, $Ax = b$, $A \in \mathbb{M}^d(\mathbb{R})$, $b \in \mathbb{R}^d$:

$$\begin{cases} x_{m+1} = Cx_m + w, & m \in \mathbb{N}, \\ C = I - M^{-1}A, & w = M^{-1}b, \qquad \det M \neq 0, \\ x_0 = \xi_0 \in \mathbb{R}^d \end{cases}$$

$$\Psi : \mathbb{R}^d \to (\mathbb{R}^d)^{\mathbb{N}}, \qquad \phi : \mathbb{R}^d \to \mathbb{R}^d, \qquad \phi(x) = Cx + w$$

O método tem o.c. 1.

Equações às diferenças

Definição. Por uma **equação às diferenças de ordem** $p \in \mathbb{N}_1$ linear e de coeficentes constantes, entende-se uma equação da forma

$$x_{m+p} + \sum_{i=0}^{p-1} a_i x_{m+i} = 0,$$

onde $m \in \mathbb{N}, x_m \in \mathbb{C}$ e os coeficientes $a_0, a_1, \ldots, a_{p-1}$ são números complexos.

Definição. O **problema de valor inicial** para a equação às diferenças consiste em determinar a solução da equação que satisfaça às condições iniciais

$$x_i = \xi_i, \qquad \forall i \in \{0, \ldots, p-1\},$$

onde $\xi_0, \xi_1, \ldots, \xi_{p-1}$ são dados. Este problema constitui um método iterativo a p passos de acordo com a definição apresentada anteriormente.

Exemplo. A sucessão gerada pelo método iterativo

$$\begin{cases} x_{m+2} - x_{m+1} - x_m = 0, & m \geq 0, \\ x_0 = x_1 = 1, \end{cases}$$

é conhecida como **sucessão de Fibonacci.**

Exemplo. Solução do problema de valor inicial

$$\begin{cases} x_{m+2} + a_1 x_{m+1} + a_0 x_m = 0, & m \geq 0, \\ x_0 = \xi_0, \quad x_1 = \xi_1. \end{cases}$$

Sendo r_1, r_2 os zeros do polinómio $P(r) = r^2 + a_1 r + a_0$ tem-se:

(1) $r_1 \neq r_2$

$$x_m = c_1 r_1^m + c_2 r_2^m, \qquad c_1 = \frac{r_2\xi_0 - \xi_1}{r_2 - r_1}, \qquad c_2 = \frac{-r_1\xi_0 + \xi_1}{r_2 - r_1}$$

(2) $r_1 = r_2$

$$x_m = r_1^m(c_1 + c_2 m), \qquad c_1 = \xi_0, \qquad c_2 = \frac{\xi_1}{r_1} - \xi_0$$

Este resultado pode obter-se escrevendo a solução (1) na forma

$$x_m = \xi_0 r_1^m + (\xi_1 - r_1\xi_0)\,\frac{r_2^m - r_1^m}{r_2 - r_1},$$

e fazendo $r_2 \to r_1$.

Proposição. Considere-se a equação às diferenças de ordem p

$$x_{m+p} + \sum_{i=0}^{p-1} a_i x_{m+i} = 0.$$

Seja

$$P(z) = z^p + \sum_{i=0}^{p-1} a_i z^i,$$

o *polinómio característico* da equação. Então:

(1) (i) Se P tem p raízes distintas $z_1, z_2, \ldots, z_p$, a solução geral da equação é

$$x_m = \sum_{i=1}^{p} C_{i0} z_i^m.$$

(ii) Se P tem q raízes distintas, $1 \leq q < p$, $z_1, \ldots, z_q$, com multiplicidades $\mu_1, \ldots, \mu_q$, respectivamente, verificando a identidade $\mu_1 + \ldots + \mu_q = p$, a solução geral da equação é

$$x_m = \sum_{i=1}^{q} \left(\sum_{j=0}^{\mu_i - 1} C_{ij}\, m^j \right) z_i^m.$$

Em ambos os casos os C_{ij} são constantes arbitrárias.

(2) O problema de valor inicial tem uma solução única, com as constantes C_{ij} fixadas pelas condições iniciais.

Exemplo. A sucessão de Fibonacci tem o termo geral

$$x_m = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^{m+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{m+1} \right].$$

Com efeito:

Polinómio característico: $P(z) = z^2 - z - 1$

Raízes de P: $r_1 = \dfrac{1+\sqrt{5}}{2}, \quad r_2 = \dfrac{1-\sqrt{5}}{2}$

Solução geral: $x_m = C_1 r_1^m + C_2 r_2^m$

Condições iniciais: $\begin{cases} 1 = x_0 = C_1 + C_2 \\ 1 = x_1 = C_1 r_1 + C_2 r_2 \end{cases} \Rightarrow \begin{cases} C_1 = \dfrac{r_1}{\sqrt{5}} \\ C_2 = -\dfrac{r_2}{\sqrt{5}} \end{cases}$

Exemplo. **Instabilidade numérica** ilustrada pelo método iterativo a dois passos definido por

$$\begin{cases} x_{m+2} = \dfrac{13}{3}\, x_{m+1} - \dfrac{4}{3}\, x_m, & m \geq 0, \\ x_0 = 1, \;\; x_1 = \dfrac{1}{3}, \end{cases}$$

cuja solução exacta é

$$x_m = \left(\frac{1}{3} \right)^m.$$

Na página seguinte apresentam-se os resultados do cálculo numérico da sucessão $\tilde{x}_m$ gerada pelo método iterativo em precisão simples e dupla.

Os resultados podem ser interpretados a partir da solução exacta do problema de valor inicial perturbado,

$$\begin{cases} x_{m+2} = \dfrac{13}{3}\, x_{m+1} - \dfrac{4}{3}\, x_m, & m \geq 0, \\ x_0 = 1 + \varepsilon_0, \;\; x_1 = \dfrac{1}{3} + \varepsilon_1, \end{cases}$$

onde $|\varepsilon_0|, |\varepsilon_1| \ll 1$:

$$x_m = \left[1 + \frac{3}{11} (4\varepsilon_0 - \varepsilon_1) \right] \left(\frac{1}{3} \right)^m + \frac{1}{11} (3\varepsilon_1 - \varepsilon_0)(4)^m.$$

Com efeito:

Polinómio característico: $P(z) = z^2 - \frac{13}{3}z + \frac{4}{3}$

Raízes de P: $r_1 = 4, \qquad r_2 = \frac{1}{3}$

Solução geral: $x_m = C_1 r_1^m + C_2 r_2^m$

Condições iniciais: $\begin{cases} 1 + \varepsilon_0 = x_0 = C_1 + C_2 \\ \frac{1}{3} + \varepsilon_1 = x_1 = C_1 r_1 + C_2 r_2 \end{cases} \Rightarrow \begin{cases} C_1 = \frac{1}{11}(3\varepsilon_1 - \varepsilon_0) \\ C_2 = 1 + \frac{3}{11}(4\varepsilon_0 - \varepsilon_1) \end{cases}$

Cálculo da sucessão $\t{x}_m$ em precisão simples:

m	$\t{x}_m$	x_m	$1 - \t{x}_m/x_m$
0	1.	1.	0.
1	0.333333343	0.333333343	0.
2	0.111111164	0.111111119	-4.02331324E-07
3	0.0370372571	0.037037041	-5.83380415E-06
4	0.0123465639	0.012345681	-7.15143833E-05
5	0.00411876757	0.00411522714	-0.000860322558
6	0.00138590776	0.00137174246	-0.010326501
7	0.000513910258	0.000457247486	-0.123921447
8	0.000379067467	0.000152415843	-1.48706079
9	0.000957412063	5.08052835E-05	-17.8447342
10	0.00364336255	1.69350951E-05	-214.136826
11	0.0145113552	5.64503171E-06	-2569.64185
12	0.0580247231	1.88167735E-06	-30835.7012
13	0.232092008	6.27225802E-07	-370028.438
14	0.928365767	2.09075282E-07	-4440341.
15	3.71346235	6.96917581E-08	-53284096.

Cálculo da sucessão $\t{x}_m$ em precisão dupla:

m	$\t{x}_m$	x_m	$1 - \t{x}_m/x_m$
0	1.	1.	0.
1	0.333333333	0.333333333	9.99200722E-15
2	0.111111111	0.111111111	1.30770395E-13
3	0.037037037	0.037037037	1.58029839E-12
4	0.012345679	0.012345679	1.89748217E-11
5	0.00411522634	0.00411522634	2.27709242E-10
6	0.00137174211	0.00137174211	2.73252339E-09
7	0.000457247356	0.000457247371	3.27902927E-08
8	0.00015241573	0.00015241579	3.93483525E-07
9	5.08050235E-05	5.08052634E-05	4.72180232E-06
10	1.69341282E-05	1.69350878E-05	5.66616278E-05
11	5.64119099E-06	5.64502927E-06	0.000679939534
12	1.86632331E-06	1.88167642E-06	0.00815927441
13	5.65813017E-07	6.27225474E-07	0.0979112929
14	-3.65746704E-08	2.09075158E-07	1.17493551
15	-9.12907595E-07	6.96917194E-08	14.0992262

3. RESOLUÇÃO NUMÉRICA DE EQUAÇÕES NÃO-LINEARES

Localização de raízes: teoremas elementares da Análise Matemática

• Sendo $f \in C([a,b])$ a equação $f(x) = 0$ pode não ter soluções, pode ter uma única solução ou pode ter mais do que uma solução no intervalo $[a,b]$.

Exemplo. A função $f_\varepsilon : [-3,3] \to \mathbb{R}, \quad f_\varepsilon(x) = x^4 - 4x^2 + \varepsilon$, tem um número de zeros no intervalo $[-3,3]$ que varia com o parâmetro ε entre zero e quatro.

Valores de ε	$]-\infty,-45[$	$[-45,0[$	0	$]0,4[$	4	$]4,\infty[$
Número de zeros de f_ε	0	2	3	4	2	0

• Os resultados seguintes (*Introdução à Análise Matemática*, Jaime Campos Ferreira, Fundação Calouste Gulbenkian) são essenciais para a localização de raízes de funções reais de variável real.

Teorema (do valor intermédio). Seja f uma função contínua num intervalo I, a e b dois pontos de I tais que $f(a) \neq f(b)$; então, qualquer que seja o númro k (estritamente) compreendido entre $f(a)$ e $f(b)$ existe pelo menos um ponto c (estritamente) compreendido entre a e b tal que $f(c) = k$.

Teorema (de Rolle). Seja f uma função contínua no intervalo $[a,b]$ (com $a,b \in \mathbb{R},\ a < b$) e diferenciável em $]a,b[$; se $f(a) = f(b)$, existe um ponto $c \in]a,b[$ tal que $f'(c) = 0$.

Corolário. Entre dois zeros de uma função diferenciável num intervalo há, pelo menos, um zero da sua derivada.

Corolário. Entre dois zeros consecutivos da derivada de uma função diferenciável num intervalo não pode haver mais de um zero dessa função.

Teorema (de Lagrange). Se f é uma função contínua no intervalo $[a,b]$ e diferenciável em $]a,b[$, existe pelo menos um ponto $c \in]a,b[$ tal que

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

• O seguinte corolário do teorema de Lagrange é útil na obtenção de estimativas de erros de aproximações de zeros de funções:

Corolário. Seja $\tilde{z}$ uma aproximação do zero z da função $f \in C^1([\tilde{z}; z])$. Então:

$$|e_{\tilde{z}}| = |z - \tilde{z}| \leq \frac{|f(\tilde{z})|}{\min\limits_{\xi \in [\tilde{z};z]} |f'(\xi)|}.$$

Localização de raízes: método da bissecção

Definição. Seja $f \in C([a,b])$ tal que $f(a)f(b) < 0$, pelo que f tem pelo menos um zero no intervalo $[a,b] = I$. O **método da bissecção** consiste em construir a partir do intervalo $I_0 = [a_0, b_0] = I$ uma sucessão de subintervalos $I_m = [a_m, b_m] \subset I$, $m \in \mathbb{N}_1$, em que um dos extremos de I_m é o ponto médio de I_{m-1},

$$x_{m-1} = \frac{1}{2}(a_{m-1} + b_{m-1}),$$

e o outro é ou a_{m-1} ou b_{m-1}, por forma a que $f(a_m)f(b_m) < 0$, o que se consegue pondo:

(i) $a_m = a_{m-1}$, $b_m = x_{m-1}$, se $f(a_{m-1})f(x_{m-1}) < 0$;

(ii) $a_m = x_{m-1}$, $b_m = b_{m-1}$, se $f(x_{m-1})f(b_{m-1}) < 0$.

Notando que

$$b_m - a_m = \frac{1}{2}(b_{m-1} - a_{m-1}),$$

verifica-se que um zero de f vai sendo sucessivamente confinado a intervalos cada vez mais pequenos.

Nota. O método da bissecção pode descrever-se analíticamente pela fórmula:

$$x_{m+1} = x_m + \frac{b-a}{2^{m+2}} \operatorname{sgn}[f(a)f(x_m)], \quad m \in \mathbb{N}, \qquad x_0 = \frac{a+b}{2}.$$

Proposição. Seja $f \in C([a,b])$ tal que $f(a)f(b) < 0$ e que tem apenas uma raiz z em $[a,b]$. Então o método da bissecção converge para z e verificam-se as seguintes estimativas de erro:

(1) $|z - x_m| \leq \dfrac{b-a}{2^{m+1}}$, $m \geq 0$; (Estimativa "a priori")

(2) $|z - x_m| \leq |x_m - x_{m-1}|$, $m \geq 1$. (Estimativa "a posteriori")

Dem.: $(\cdots)$

Exemplo. Determinação da raiz positiva da equação $f(x) = x^2 - 2 = 0$ com um erro inferior a 10^{-4}.

A equação tem uma e uma só raiz no intervalo $I = [0,2]$ pois

$$f(0) = -2, \qquad f(2) = 2, \qquad f'(x) = 2x \geq 0, \ \ \forall x \in I.$$

Calculemos o menor valor de m para o qual $|z - x_m| \leq \varepsilon$:

$$|z - x_m| \leq \frac{b-a}{2^{m+1}} \leq \varepsilon \ \ \Rightarrow \ \ m \geq \frac{\log \dfrac{b-a}{\varepsilon}}{\log 2} - 1$$

Para $[a,b] = [0,2]$, $\varepsilon = 10^{-4}$ obtém-se $m \geq 13.3$. logo $m = 14$.

m	a_m	b_m	x_m	$f(x_m)$	$x_m - x_{m-1}$	$z - x_m$
0	0.000000	2.000000	1.000000	-1.000000		0.414214
1	1.000000	2.000000	1.500000	0.250000	0.500000	-0.085787
2	1.000000	1.500000	1.250000	-0.437500	-0.250000	0.164214
3	1.250000	1.500000	1.375000	-0.109375	0.125000	0.039214
4	1.375000	1.500000	1.437500	0.066406	0.062500	-0.023287
5	1.375000	1.437500	1.406250	-0.022461	-0.003125	0.007964
6	1.406250	1.437500	1.421875	0.021729	0.015625	-0.007662
7	1.406250	1.421875	1.414063	-0.000427	-0.007812	0.000151
8	1.414063	1.421875	1.417969	0.010635	0.003906	-0.003756
9	1.414063	1.417969	1.416016	0.005100	-0.001953	-0.001803
10	1.414063	1.416016	1.415039	0.002336	-0.000977	-0.000826
11	1.414063	1.415039	1.414551	0.000954	-0.000488	-0.000337
12	1.414063	1.414551	1.414307	0.000263	-0.000244	-0.000093
13	1.414063	1.414307	1.414185	-0.000082	-0.000122	0.000029
14	1.414185	1.414307	1.414246	0.000091	0.000061	-0.000032

Notas.

(a) O método da bissecção é um método iterativo a um passo com função iteradora descontínua.

(b) O método é sempre convergente desde que $f(a)f(b) < 0$, mas a convergência pode ser muito lenta.

(c) Não se pode afirmar que o método tem convergência linear mas apenas que tem semelhanças com métodos com convergência linear. Com efeito a sucessão dos majorantes do erro $\{w_m\}$, $w_m = \frac{b-a}{2^{m+1}}$, converge linearmente com factor assimptótico de convergência $K_\infty^{[1]} = \frac{1}{2}$.

(d) A estimativa "a posteriori" pode ser utilizada como **critério de paragem** para o método iterativo.

(e) O método da bissecção é útil para a localização de raízes e para a inicialização de métodos mais rápidos cuja convergência só é garantida com uma boa aproximação inicial.

Método do ponto fixo

Definição. Diz-se que z é um **ponto fixo** de uma função $g : \mathbb{R} \to \mathbb{R}$ se e só se $z = g(z)$.

Notas.

(a) Podemos transformar qualquer equação $f(x) = 0$, $f : \mathbb{R} \to \mathbb{R}$, numa equação $x = g(x)$, estabelecendo a equivalência

$$f(x) = 0 \quad \Leftrightarrow \quad x = g(x),$$

num certo domínio D. É claro que se $z \in D$ for zero da função f então será ponto fixo de g e vice-versa.

(b) Há infinitas possibilidades de escolher g (a função iteradora) de forma a que a equivalência se verifique. Por exemplo, basta pensar que se $\omega \neq 0$ temos

$$f(x) = 0 \quad \Leftrightarrow \quad x = x + \omega f(x),$$

Como iremos ver, algumas escolhas de g não serão apropriadas, ou serão menos apropriadas do que outras, para o objectivo em vista.

Definição. O **método do ponto fixo** é o método iterativo a um passo da forma

$$x_{m+1} = g(x_m), \quad m \in \mathbb{N}, \qquad x_0 = \xi_0 \text{ dado.}$$

Nota. Considerando g uma função contínua, se o método convergir, convergirá para um ponto fixo z que, pela equivalência estabelecida, será uma raiz da equação, ou seja, $f(z) = 0$.

Exemplo. Utilização do método do ponto fixo com as funções iteradoras:

$$\text{(a)} \;\; g(x) = x + \omega(x^2 - 3); \qquad \text{(b)} \;\; g(x) = \frac{3}{x}; \qquad \text{(c)} \;\; g(x) = \frac{1}{2}\left(x + \frac{3}{x}\right);$$

no cálculo da raiz positiva da equação $x^2 - 3 = 0$.

n	x_n			
	(a) $\omega = 1$	(b)	(c)	(a) $\omega = -1/4$
0	2.0	2.0	2.0	2.0
1	3.0	1.5	1.750000000000000	1.750000000000000
2	9.0	2.0	1.732142857142857	1.734375000000000
3	87.0	1.5	1.732050810014728	1.732360839843750
4	7563.0	2.0	1.732050807568877	1.732092319987714
5		1.5	1.732050807568877	1.732056368747609
6				1.732051552617821
7				1.732050907386370
8				1.732050820941883
9				1.732050809360520
10				1.732050807808912
11				1.732050807601036
12				1.732050807573186
13				1.732050807569455
14				1.732050807568955
15				1.732050807568888
16				1.732050807568879
17				1.732050807568877

Exemplo.

(a) Ilustração da convergência *monótona* para o ponto fixo da sucessão gerada pelo método do ponto fixo com função iteradora $g(x) = 2 - \cos x$ e valor inicial $x_0 = 3$.

(b) Ilustração da convergência *alternada* para o ponto fixo da sucessão gerada pelo método do ponto fixo com função iteradora $g(x) = 2 + 0.8\cos x$ e valor inicial $x_0 = 3$.

(c) Ilustração da não convergência para o ponto fixo da sucessão gerada pelo método do ponto fixo com função iteradora $g(x) = \frac{1}{2}(-1 + 3x - 3\sin x)$ e valor inicial $x_0 = 1.6$.

(d) Ilustração da não convergência para o ponto fixo da sucessão gerada pelo método do ponto fixo com função iteradora $g(x) = 2 + 1.5\cos x$ e valor inicial $x_0 = 1.6$.

Definição. Uma função $g : I \to \mathbb{R}$, onde I é um intervalo fechado, diz-se uma **função de Lipschitz** ou **Lipschitziana** em I se existir um número real $L \geq 0$ tal que

$$|g(x) - g(y)| \leq L|x - y|, \qquad \forall x, y \in I.$$

Se $L < 1$ a função diz-se **contractiva** ou uma **contracção** em I, com **constante de contractividade** L.

Exemplo.

(a) $f : [-a, a] \to \mathbb{R}$, $a > 0$, $f(x) = |x|$.

É função de Lipschitz com $L = 1$.

(b) $f : [a, b] \to \mathbb{R}$, $b > a > 0$, $f(x) = x^{1/2}$.

É função de Lipschitz com $L = \dfrac{1}{2\sqrt{a}}$. É contracção para $a > \dfrac{1}{4}$.

Teorema do ponto fixo (em $\mathbb{R}$). Seja $g : I \to \mathbb{R}$ uma função contractiva num intervalo $I \subset \mathbb{R}$ com constante de contractividade L e tal que $g(I) \subset I$. Então:

(1) g tem um e um só ponto fixo z em I;

(2) a sucessão $\{x_m\}_{m\in\mathbb{N}}$ definida por $x_{m+1} = g(x_m)$ converge para z, qualquer que seja $x_0 \in I$;

(3) verificam-se as seguintes estimativas de erro:

(i) $|z - x_{m+1}| \leq L|z - x_m|$;

(ii) $|z - x_m| \leq L^m|z - x_0|$;

(iii) $|z - x_m| \leq \dfrac{1}{1 - L}|x_{m+1} - x_m|$;

(iv) $|z - x_{m+1}| \leq \dfrac{L}{1 - L}|x_{m+1} - x_m|$;

(v) $|z - x_m| \leq \dfrac{L^m}{1-L}|x_1 - x_0|$.

Dem.: $(\cdots)$

Proposição. A função $g \in C^1(I)$ é contractiva em I se e só se

$$\max_{x \in I} |g'(x)| < 1.$$

Dem.: $(\cdots)$

Proposição. Seja $g \in C^1(I)$ uma função tal que:

$$\text{(i) } g(I) \subset I; \qquad \text{(ii) } L = \max_{x \in I} |g'(x)| < 1.$$

Então são válidas as conclusões do teorema do ponto fixo.

Dem.: $(\cdots)$

Proposição (da convergência monótona e alternada). Seja $g \in C^1(I)$ uma função tal que $g(I) \subset I$.

(1) Seja $0 < g'(x) < 1, \ \ \forall x \in I$. Se $x_0 \in I$ há convergência monótona do método do ponto fixo (isto é, as iteradas ficam todas à esquerda (respectivamente à direita) do ponto fixo se a iterada inicial estiver à esquerda (respectivamente à direita) do ponto fixo).

(2) Seja $-1 < g'(x) < 0, \ \ \forall x \in I$. Se $x_0 \in I$ há convergência alternada do método do ponto fixo (isto é, as iteradas ficam alternadamente à esquerda e à direita do ponto fixo).

Dem.: $(\cdots)$

Proposição (da convergência local). Seja $g \in C^1(V_z)$, onde V_z é uma vizinhança de z, ponto fixo de g, com $|g'(z)| < 1$. Então são válidas as conclusões do teorema do ponto fixo desde que x_0 esteja suficientemente perto de z.

Dem.: $(\cdots)$

Proposição (da não convergência para um ponto fixo). Seja $g \in C^1(V_z)$, onde V_z é uma vizinhança de z, ponto fixo de g, com $|g'(z)| > 1$. Então a sucessão $\{x_m\}, x_{m+1} = g(x_m)$, não pode convergir para esse ponto fixo (a não ser que "excepcionalmente" $x_m = z$ para algum m).

Dem.: $(\cdots)$

Proposição (da convergência local, linear). Seja $g \in C^1(V_z)$, onde V_z é uma vizinhança de

z, ponto fixo de g, com $0 < |g'(z)| < 1$. Então, para qualquer x_0 para o qual o método do ponto fixo converge para z, tem-se:

(1) $z - x_{m+1} = g'(\xi_m)(z - x_m), \qquad \xi_m \in]z; x_m[$;

(2) $\lim_{m\to\infty} \dfrac{z - x_{m+1}}{z - x_m} = g'(z).$

Dem.: $(\cdots)$

Nota. O método do ponto fixo tem neste caso convergência linear com coeficiente assimptótico de convergência

$$K_\infty^{[1]} = |g'(z)| .$$

Proposição (da convergência local, supra-linear). Seja $g \in C^p(V_z), \ \ p \geq 2$, onde V_z é uma vizinhança de z, ponto fixo de g, tal que

$$g'(z) = \cdots = g^{(p-1)}(z) = 0, \qquad g^{(p)}(z) \neq 0.$$

Então, para qualquer x_0 para o qual o método do ponto fixo converge para z, tem-se:

(1) $z - x_{m+1} = \dfrac{(-1)^{p+1}}{p!}\, g^{(p)}(\xi_m)(z - x_m)^p, \qquad \xi_m \in]z; x_m[$;

(2) $\lim_{m\to\infty} \dfrac{z - x_{m+1}}{(z - x_m)^p} = \dfrac{(-1)^{p+1}}{p!}\, g^{(p)}(z)$.

Dem.: $(\cdots)$

Nota. O método do ponto fixo tem neste caso convergência de ordem p com coeficiente assimptótico de convergência

$$K_\infty^{[p]} = \frac{1}{p!} \left| g^{(p)}(z) \right| .$$

Exemplo. Considere-se o polinómio do 3º grau

$$p(x) = x^3 - 4x + 1.$$

(a) Mostrar que o polinómio tem três raízes reais $z_1 < z_2 < z_3$ tais que

$$z_1 \in [-2.2, -2.0], \qquad z_2 \in [0.1, 0.3], \qquad z_3 \in [1.8, 2.0].$$

(b) Mostrar que o método do ponto fixo com função iteradora g_1, definida por

$$g_1(x) = (4x - 1)^{1/3}, \qquad x \in \mathbb{R},$$

converge para a raiz z_1 para qualquer $x_0 \in [-2.2, -2.0]$.

(c) Utilizar o método do ponto fixo com função iteradora g_1 para obter um valor aproximado da raiz z_1 com um erro absoluto inferior a 10^{-6}.

(d) Mostrar que o método do ponto fixo com função iteradora g_1 pode ser utilizado para calcular a raiz z_3 mas não a raiz z_2.

(e) Mostrar que o método do ponto fixo com função iteradora g_2, definida por

$$g_2(x) = \frac{1}{4}\,(x^3 + 1), \qquad x \in \mathbb{R},$$

pode ser utilizado para calcular a raiz z_2 mas não qualquer das outras raízes.

(f) Mostrar que o método do ponto fixo com função iteradora g_3, definida por

$$g_3(x) = \frac{4}{x} - \frac{1}{x^2} \qquad x \neq 0,$$

pode ser utilizado para calcular a raiz z_3 mas não qualquer das outras raízes.

Resolução:

(a) $\left.\begin{array}{l} p(-2.2) = -0.848 \\ p(-2.0) = 1.0 \end{array}\right\} \Rightarrow$ p tem pelo menos uma raiz em $I_1 = [-2.2, -2.0]$

$\left.\begin{array}{l} p(0.1) = 0.601 \\ p(0.3) = -0.173 \end{array}\right\} \Rightarrow$ p tem pelo menos uma raiz em $I_2 = [0.1, 0.3]$

$\left.\begin{array}{l} p(1.8) = -0.368 \\ p(2.0) = 1.00 \end{array}\right\} \Rightarrow$ p tem pelo menos uma raiz em $I_3 = [1.8, 2.0]$

Sendo p um polinómio do 3^o grau conclui-se que p tem exactamente três raízes reais, uma em cada um dos intervalos indicados.

(b) $p(x) = 0 \;\Leftrightarrow\; x = g_1(x)$

$$g_1(x) = (4x-1)^{1/3}, \quad g_1'(x) = \frac{4}{3}\,(4x-1)^{-2/3}, \quad g_1''(x) = -\frac{32}{9}\,(4x-1)^{-5/3}$$

$\left.\begin{array}{l} g_1(-2.2) = -2.14 \in I_1 \\ g_1(-2.0) = -2.08 \in I_1 \\ g_1'(x) > 0, \;\; \forall x \in I_1 \end{array}\right\} \Rightarrow$ $g_1(I_1) \subset I_1$

$\left.\begin{array}{l} g_1'(x) > 0, \;\; \forall x \in I_1 \\ g_1''(x) > 0, \;\; \forall x \in I_1 \end{array}\right\} \Rightarrow \max\limits_{x \in I_1} |g_1'(x)| = |g_1'(-2.0)| = 0.308161 < 1$

$\Rightarrow$ g_1 tem um e um só ponto fixo $z_1 \in I_1$ para o qual o método do ponto fixo com função iteradora g_1 converge, $\forall x_0 \in I_1$.

(c) $x_{m+1} = g_1(x_m), \;\; m \in \mathbb{N}, \qquad x_0 \in I_1$

$$|z_1 - x_m| \leq \frac{L}{1-L}\ |x_m - x_{m-1}| =: B_m$$

$$L = \max_{x \in I_1} |g_1'(x)| = 0.308161$$

m	x_m	B_m
0	-2.1	
1	-2.11045429	0.466×10^{-2}
2	-2.11357921	0.139×10^{-2}
3	-2.11451150	0.415×10^{-3}
4	-2.11478948	0.124×10^{-3}
5	-2.11487235	0.369×10^{-4}
6	-2.11489705	0.110×10^{-4}
7	-2.11490441	0.328×10^{-5}
8	-2.11490661	0.978×10^{-6}

$z_1 \approx x_8 = -2.11491$

Note-se que $z_1 = -2.114907541476755\ldots$

$$|z_1 - x_7| = 0.313 \times 10^{-5}, \qquad |z_1 - x_8| = 0.932 \times 10^{-6}$$

(d) $\left.\begin{array}{l} g_1'(x) > 0, \ \forall x \in I_3 \\ g_1''(x) < 0, \ \forall x \in I_3 \end{array}\right\} \Rightarrow \max_{x \in I_3} |g_1'(x)| = |g_1'(1.8)| = 0.395 < 1$

$\left.\begin{array}{l} g_1(1.8) = 1.84 \in I_3 \\ g_1(2.0) = 1.91 \in I_3 \\ g_1'(x) > 0, \ \forall x \in I_3 \end{array}\right\} \Rightarrow g_1(I_3) \subset I_3.$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_1 converge para z_3, $\forall x_0 \in I_3$.

$\tilde{I}_2 = [0.251, 0.3]$

$\left.\begin{array}{l} p(0.251) = 0.0118 \\ p(0.3) = -0.173 \end{array}\right\} \Rightarrow z_2 \in \tilde{I}_2$

$\left.\begin{array}{l} g_1'(x) > 0, \ \forall x \in \tilde{I}_2 \\ g_1''(x) < 0, \ \forall x \in \tilde{I}_2 \end{array}\right\} \Rightarrow \min_{x \in \tilde{I}_2} |g_1'(x)| = |g_1'(0.3)| = 3.90 > 1$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_1 não pode convergir para z_2.

(e) $g_2(x) = \frac{1}{4}(x^3 + 1), \qquad g_2'(x) = \frac{3x^2}{4}, \qquad g_2''(x) = \frac{3x}{2}$

$\left.\begin{array}{l} g_2'(x) > 0, \ \forall x \in I_2 \\ g_2''(x) > 0, \ \forall x \in I_2 \end{array}\right\} \Rightarrow \max_{x \in I_2} |g_2'(x)| = |g_2'(0.3)| = 0.0675 < 1$

$$\left.\begin{array}{l} g_2(0.1) = 0.250 \in I_2 \\ g_2(0.3) = 0.257 \in I_2 \\ g_2'(x) > 0, \ \ \forall x \in I_2 \end{array}\right\} \ \Rightarrow \ \ g_2(I_2) \subset I_2.$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_2 converge para z_2, $\forall x_0 \in I_2$.

$$\left.\begin{array}{l} g_2'(x) > 0, \ \ \forall x \in I_1 \\ g_2''(x) < 0, \ \ \forall x \in I_1 \end{array}\right\} \ \Rightarrow \ \ \min_{x \in I_1} |g_2'(x)| = |g_2'(-2.0)| = 3.0 > 1$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_2 não pode convergir para z_1.

$$\left.\begin{array}{l} g_2'(x) > 0, \ \ \forall x \in I_3 \\ g_2''(x) > 0, \ \ \forall x \in I_3 \end{array}\right\} \ \Rightarrow \ \ \min_{x \in I_3} |g_2'(x)| = |g_2'(1.8)| = 2.43 > 1$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_2 não pode convergir para z_3.

(f) $g_3(x) = \dfrac{4}{x^2}\left(x - \dfrac{1}{4}\right), \qquad g_3'(x) = \dfrac{4}{x^3}\left(\dfrac{1}{2} - x\right), \qquad g_3''(x) = \dfrac{8}{x^4}\left(x - \dfrac{3}{4}\right)$

$$\left.\begin{array}{l} g_3'(x) < 0, \ \ \forall x \in I_3 \\ g_3''(x) > 0, \ \ \forall x \in I_3 \end{array}\right\} \ \Rightarrow \ \ \max_{x \in I_3} |g_3'(x)| = |g_3'(1.8)| = 0.892 < 1$$

$$\left.\begin{array}{l} g_3(1.8) = 1.91 \in I_3 \\ g_3(2.0) = 1.75 \notin I_3 \end{array}\right\} \ \Rightarrow \ \ g_3(I_3) \not\subset I_3$$

$\tilde{I}_3 = [1.8, 1.93]$

$$\left.\begin{array}{l} p(1.8) = -0.368 \\ p(1.93) = 0.469 \end{array}\right\} \ \Rightarrow \ \ z_3 \in \tilde{I}_3$$

$$\left.\begin{array}{l} g_3(1.8) = 1.91358 \in \tilde{I}_3 \\ g_3(1.93) = 1.80408 \in \tilde{I}_3 \\ g_3'(x) < 0, \ \ \forall x \in \tilde{I}_3 \end{array}\right\} \ \Rightarrow \ \ g_3(\tilde{I}_3) \subset \tilde{I}_3.$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_3 converge para z_3, $\forall x_0 \in \tilde{I}_3$.

$$\left.\begin{array}{l} g_3'(x) < 0, \ \ \forall x \in I_1 \\ g_3''(x) < 0, \ \ \forall x \in I_1 \end{array}\right\} \ \Rightarrow \ \ \min_{x \in I_1} |g_3'(x)| = |g_3'(-2.2)| = 1.01 > 1$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_3 não pode convergir para z_1.

$$\left.\begin{array}{l} g_3'(x) > 0, \ \ \forall x \in I_2 \\ g_3''(x) < 0, \ \ \forall x \in I_2 \end{array}\right\} \ \Rightarrow \ \ \min_{x \in I_2} |g_3'(x)| = |g_3'(0.3)| = 29.6 > 1$$

$\Rightarrow$ O método do ponto fixo com f. iteradora g_3 não pode convergir para z_2.

Método de Newton

Definição. O **método de Newton** é o método iterativo a um passo da forma

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, \quad m \in \mathbb{N}, \qquad x_0 = \xi_0 \text{ dado.}$$

Nota. O método de Newton pode ser interpretado geométricamente da seguinte forma: a iterada x_{m+1} é a intersecção do eixo das abcissas com a recta tangente à curva $y = f(x)$ no ponto $(x_m, f(x_m))$, cuja equação é

$$y = f(x_m) + f'(x_m)(x - x_m).$$

Proposição (condições suficientes de convergência). Seja $f \in C^2(I)$, $I = [a, b]$, uma função que verifica as seguintes condições:

(1) $f(a)f(b) \leq 0$;

(2) $f'(x) \neq 0, \ \forall x \in I$;

(3) $f''(x) \geq 0$ ou $f''(x) \leq 0, \ \forall x \in I$;

(4) (a) $f(x_0)f''(x) \geq 0, \ \forall x \in I, \quad x_0 \in I$;

ou

(b) $\left|\frac{f(a)}{f'(a)}\right| \leq b - a \quad$ e $\quad \left|\frac{f(b)}{f'(b)}\right| \leq b - a.$

Então o método de Newton converge monotonamente para a única solução z de $f(x) = 0$ em I.

Dem.: $(\cdots)$

Proposição (da convergência local, quadrática). Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, raiz de f, tal que $f'(z) \neq 0, \quad f''(z) \neq 0$. Então, para qualquer x_0 suficientemente próximo de z, a sucessão $\{x_m\}$ definida pelo método de Newton converge para z. Além disso:

(1) (Fórmula do erro)

$$z - x_{m+1} = -\frac{f''(\xi_m)}{2f'(x_m)} \left(z - x_m\right)^2, \qquad \xi_m \in]z; x_m[$$

(2) (Estimativa do erro)

$$|z - x_{m+1}| \leq K|z - x_m|^2, \qquad K = \frac{\max_{x \in I} |f''(x)|}{2 \min_{x \in I} |f'(x)|},$$

ou

$$|z - x_m| \leq \frac{1}{K} \left(K|z - x_0|\right)^{2^m}, \qquad m \geq 0,$$

onde $I = [z-\varepsilon, z+\varepsilon] \subset V_z$ é tal que $f'(x) \neq 0,\ \forall x \in I,\ K\varepsilon < 1$, e $x_0 \in I$.

(3) (Ordem de convergência 2)

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{(z-x_m)^2} = -\frac{f''(z)}{2f'(z)}.$$

Dem.: $(\cdots)$

Nota. O método de Newton tem pois convergência de ordem 2 com coeficiente assimptótico de convergência

$$K_\infty^{[2]} = \left|\frac{f''(z)}{2f'(z)}\right|.$$

• O método de Newton pode ser encarado como um caso particular do método do ponto fixo com função iteradora g definida por

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Usando os resultados sobre a convergência local e supra-linear do método do ponto fixo obtemos o seguinte resultado:

Proposição (da convergência local, quadrática e supra-quadrática). Seja $f \in C^{p+1}(V_z),\ p \geq 2$, onde V_z é uma vizinhança de z, raiz de f, tal que:

(i) $f'(z) \neq 0, \quad f''(z) \neq 0, \qquad$ se $p = 2$;

(ii) $f'(z) \neq 0, \quad f''(z) = \cdots = f^{(p-1)}(z) = 0, \quad f^{(p)}(z) \neq 0, \qquad$ se $p \geq 3$.

Então, para qualquer x_0 suficientemente próximo de z, o método de Newton converge para z, e

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{(z-x_m)^p} = (-1)^{p+1}\,\frac{p-1}{p!}\,\frac{f^{(p)}(z)}{f'(z)}.$$

Dem.: $(\cdots)$

Exemplo. Considere-se o polinómio do 3º grau

$$p(x) = x^3 - 4x + 1,$$

para o qual se verificou a existência de três raízes reais $z_1 < z_2 < z_3$ tais que

$$z_1 \in [-2.2, -2.0], \qquad z_2 \in [0.1, 0.3], \qquad z_3 \in [1.8, 2.0].$$

(a) Mostrar que o método de Newton com iterada inicial $x_0 \in [0.1, 0.3]$ converge para a raiz z_2.

(b) Utilizar o método de Newton para obter um valor aproximado da raiz z_2 com um erro absoluto inferior a 10^{-6}.

Resolução:

(a) Condições suficientes de convergência $\forall x_0 \in I_2 = [0.1, 0.3] = [a, b]$:

(o) $p \in C^2(I_2)$

(i) $\left.\begin{array}{l} p(a) = 0.601 \\ p(b) = -0.173 \end{array}\right\} \Rightarrow \quad p(a)p(b) < 0$

(ii) $p'(x) = 3x^2 - 4 < 0, \quad \forall x \in I_2$

(iii) $p''(x) = 6x > 0, \quad \forall x \in I_2$

(iv) $\left|\dfrac{p(a)}{p'(a)}\right| = \dfrac{0.601}{3.97} = 0.151 < 0.2 = b - a$

$\left|\dfrac{p(b)}{p'(b)}\right| = \dfrac{0.173}{3.73} = 0.0464 < 0.2 = b - a$

(b) $x_{m+1} = g(x_m), \quad m \in \mathbb{N}, \qquad x_0 \in I_2, \qquad g(x) = x - \dfrac{p(x)}{p'(x)}$

$|e_m| \leq K\,|e_{m-1}|^2 =: B_m, \quad m \geq 1$

$K = \dfrac{\max_{x\in I_2} |p''(x)|}{2\min_{x\in I_2} |p'(x)|} = \dfrac{|p''(b)|}{2|p'(b)|} = \dfrac{1.8}{2 \times 3.73} = 0.241287 \approx 0.242$

$x_0 = \dfrac{a+b}{2} = 0.2, \qquad |e_0| \leq \dfrac{b-a}{2} = 0.1$

$K|e_0| \leq 0.0242 < 1$

m	x_m	B_m
0	0.2	
1	0.2536082474226804	0.242×10^{-2}
2	0.2541016396741136	0.142×10^{-5}
3	0.2541916883650519	0.486×10^{-12}

$z_2 \approx x_3 = 0.254191688365$

Note-se que $z_2 = 0.2541916883650524\ldots, \qquad |z_2 - x_3| = 0.5 \times 10^{-15}$

Definição. Diz-se que uma função f tem um zero z de multiplicidade $\mu > 1$ se existir uma função h contínua em z tal que $h(z) \neq 0$ e que

$$f(x) = (x - z)^{\mu} h(x).$$

Nota. Se $h \in C^{\mu}(V_z)$ então

$$f(z) = f'(z) = \cdots = f^{(\mu-1)}(z) = 0, \qquad f^{(\mu)}(z) \neq 0.$$

Proposição (da convergência local, linear). Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, zero de multiplicidade $\mu > 1$ de f. Então, para qualquer x_0 suficientemente próximo de z, o método de Newton converge para z e

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{z - x_m} = 1 - \frac{1}{\mu}.$$

Dem.: $(\cdots)$

Definição. O método de Newton μ-modificado para um zero de multiplicidade $\mu > 1$ de f é definido por

$$x_{m+1} = x_m - \mu \frac{f(x_m)}{f'(x_m)}, \quad m \in \mathbb{N}, \qquad x_0 = \xi_0 \text{ dado.}$$

Proposição (da convergência local, quadrática). Seja $f \in C^3(V_z)$, onde V_z é uma vizinhança de z, zero de multiplicidade $\mu > 1$ de f. Então, para qualquer x_0 suficientemente próximo de z, o método de Newton μ-modificado converge para z e:

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{(z - x_m)^2} = -\frac{1}{\mu} \frac{h'(z)}{h(z)}.$$

onde h é tal que $f(x) = (x - z)^{\mu} h(x), \;\; h(z) \neq 0.$

Dem.: $(\cdots)$

Nota. O método de Newton μ-modificado tem a desvantagem de exigir o conhecimento "a priori" da multiplicidade da raiz.

Método da secante

Definição. O **método da secante** é o método iterativo a dois passos da forma

$$x_{m+2} = x_{m+1} - f(x_{m+1}) \frac{x_{m+1} - x_m}{f(x_{m+1}) - f(x_m)}, \quad m \in \mathbb{N}, \quad x_0 = \xi_0, \; x_1 = \xi_1, \;\; \xi_0, \xi_1 \text{ dados.}$$

Nota. O método da secante pode ser interpretado geométricamente da seguinte forma: a iterada x_{m+2} é a intersecção do eixo das abcissas com a recta que passa pelos pontos $(x_m, f(x_m))$ e $(x_{m+1}, f(x_{m+1}))$, cuja equação é

$$y = f(x_{m+1}) + \frac{f(x_{m+1}) - f(x_m)}{x_{m+1} - x_m} (x - x_{m+1}).$$

Proposição (condições suficientes de convergência). Seja $f \in C^2(I)$, $I = [a, b]$, uma função que verifica as seguintes condições:

(1) $f(a)f(b) \leq 0$;

(2) $f'(x) \neq 0, \ \forall x \in I$;

(3) $f''(x) \geq 0$ ou $f''(x) \leq 0, \ \forall x \in I$;

(4) (a) $f(x_0)f''(x) \geq 0$, e $f(x_1)f''(x) \geq 0, \ \forall x \in I, \ x_0, x_1 \in I$;
ou
(b) $\left|\dfrac{f(a)}{f'(a)}\right| \leq b - a \quad$ e $\quad \left|\dfrac{f(b)}{f'(b)}\right| \leq b - a.$

Então o método da secante converge para a única solução z de $f(x) = 0$ em I.

Proposição (da convergência local). Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, raiz de f, tal que $|f'(z)| \neq 0$. Então, para quaisquer x_0 e x_1 suficientemente próximos de z, o método da secante converge para z. Além disso:

(1) (Fórmula do erro)

$$z - x_{m+1} = -\frac{f''(\eta_m)}{2f'(\xi_m)} \, (z - x_m) \, (z - x_{m-1}) \,, \qquad \xi_m \in]x_{m-1}; x_m[, \ \ \eta_m \in]x_{m-1}; z; x_m[$$

(2) (Estimativa do erro)

$$|z - x_{m+1}| \leq K \, |z - x_m| \, |z - x_{m-1}| \,, \qquad K = \frac{\max_{x \in I} |f''(x)|}{2 \min_{x \in I} |f'(x)|},$$

ou

$$|z - x_m| \leq \frac{1}{K} \, \delta^{q_m}, \qquad m \geq 0,$$

onde $I = [z - \varepsilon, z + \varepsilon] \subset V_z$ é tal que $f'(x) \neq 0, \forall x \in I, \ K\varepsilon < 1, \ x_0, x_1 \in I,$

$$\delta = \max\{K|z - x_0|, K|z - x_1|\} < 1,$$

e $\{q_m\}$ é a sucessão de Fibonnaci.

(3) (Ordem de convergência)

$$\lim_{m \to \infty} \frac{|z - x_{m+1}|}{|z - x_m|^r} = \left|\frac{f''(z)}{2f'(z)}\right|^{r-1} =: K_\infty^{[r]}, \qquad r = \frac{\sqrt{5} + 1}{2} = 1.61803\ldots$$

Dem.: $(\cdots)$

Exemplo. Considere-se o polinómio do 3^o grau

$$p(x) = x^3 - 4x + 1,$$

para o qual se verificou a existência de três raízes reais $z_1 < z_2 < z_3$ tais que

$$z_1 \in [-2.2, -2.0], \qquad z_2 \in [0.1, 0.3], \qquad z_3 \in [1.8, 2.0].$$

(a) Mostrar que o método da secante com iteradas iniciais $x_0, x_1 \in [1.8, 2.0]$ converge para a raiz z_3.

(b) Utilizar o método da secante para obter um valor aproximado da raiz z_3 com um erro absoluto inferior a 10^{-6}.

Resolução:

(a) Condições suficientes de convergência $\forall x_0 \in I_3 = [1.8, 2.0] = [a, b]$:

(o) $p \in C^2(I_3)$

(i) $\left.\begin{array}{l} p(a) = -0.368 \\ p(b) = 1.00 \end{array}\right\} \Rightarrow \; p(a)p(b) < 0$

(ii) $p'(x) = 3x^2 - 4 > 0, \;\; \forall x \in I_3$

(iii) $p''(x) = 6x > 0, \;\; \forall x \in I_3$

(iv) $\left|\dfrac{p(a)}{p'(a)}\right| = 0.0643 < 0.2 = b - a$

$\left|\dfrac{p(b)}{p'(b)}\right| = 0.125 < 0.2 = b - a$

(b) $x_{m+2} = g(x_m, x_{m-1}), \quad m \in \mathbb{N}, \qquad g(x, y) = \dfrac{xp(y) - yp(x)}{p(y) - p(x)}$

$|e_m| \leq K\,|e_{m-1}||e_{m-2}| =: B_m, \quad m \geq 2$

$K = \dfrac{\max_{x \in I_3} |p''(x)|}{2 \min_{x \in I_3} |p'(x)|} = \dfrac{|p''(b)|}{2|p'(a)|} = \dfrac{12.0}{2 \times 5.72} = 1.05$

$|e_0| \leq b - a = 0.2, \qquad |e_1| \leq b - a = 0.2$

$K|e_0| \leq 0.21 < 1, \qquad K|e_1| \leq 0.21 < 1$

m	x_m	B_m
0	1.8	
1	2.0	
2	1.853801169590643	0.420×10^{-1}
3	1.860025945055839	0.882×10^{-2}
4	1.860810653297940	0.389×10^{-3}
5	1.860805849838245	0.360×10^{-5}
6	1.860805853111690	0.147×10^{-8}

$z_3 \approx x_6 = 1.86080585$

Note-se que $z_3 = 1.860805853111703\ldots, \qquad |z_3 - x_6| = 0.140 \times 10^{-13}$

Nota (comparação do método de Newton com o método da secante). Vimos que;

- ◇ O método de Newton:
  (i) tem ordem de convergência 2;
  (ii) envolve o cálculo de f e f' em cada iteração.
- ◇ O método da secante:
  (i) tem ordem de convergência $r \approx 1.618$;
  (ii) envolve apenas o cálculo de f em cada iteração.

Pode mostrar-se que o tempo de cálculo dos dois métodos, para um mesmo erro, é o mesmo se

$$t_{f'} \approx 0.44 t_f$$

ou, dito de outra maneira, o método de Newton será mais eficaz se

$$t_{f'} < 0.44 t_f.$$

4. RESOLUÇÃO NUMÉRICA DE SISTEMAS LINEARES

Normas matriciais

• Uma vez que $\mathbb{M}^n(\mathbb{R})$ ou $\mathbb{M}^n(\mathbb{C})$ são espaços vectoriais de dimensão n^2, topologicamente equivalentes aos espaço $\mathbb{R}^{n^2}$ ou $\mathbb{C}^{n^2}$, respectivamente, qualquer das normas que introduzimos em espaços vectoriais também serve como norma matricial. No entanto nem todas as normas possíveis têm interesse prático. Vamos, portanto, começar por definir duas condições suplementares que as normas matriciais devem satisfazer para se tornarem úteis.

Definição. Seja M uma norma no espaço $\mathbb{M}^n(\mathbb{C})$. A norma M diz-se **regular** se e só se for satisfeita a condição

$$\|AB\|_M \leq \|A\|_M \|B\|_M, \qquad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$$

Definição. Seja M uma norma no espaço $\mathbb{M}^n(\mathbb{C})$ e V uma norma no espaço vectorial $\mathbb{C}^n$. A norma M diz-se **compatível** com a norma V se e só se for satisfeita a condição

$$\|Ax\|_V \leq \|A\|_M \|x\|_V, \qquad \forall A \in \mathbb{M}^n(\mathbb{C}), \quad \forall x \in \mathbb{C}^n.$$

• A definição seguinte permite-nos obter normas matriciais que satisfazem às cinco condições impostas.

Definição. Diz-se que uma norma matricial M em $\mathbb{M}^n(\mathbb{C})$ está **associada** a uma certa norma vectorial V em $\mathbb{C}^n$ se e só se ela for definida pela igualdade

$$\|A\|_M = \sup_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\|Ax\|_V}{\|x\|_V}. \qquad (\#)$$

Também se diz neste caso que a norma M é a norma **induzida** em $\mathbb{M}^n(\mathbb{C})$ pela norma vectorial V em $\mathbb{C}^n$ ou é a norma em $\mathbb{M}^n(\mathbb{C})$ **subordinada** à norma vectorial V em $\mathbb{C}^n$.

Proposição. A equação (#) define uma norma matricial, regular e compatível com a norma V, isto é, que satisfaz às cinco condições:

(1) $\|A\|_M \geq 0, \quad \forall A \in \mathbb{M}^n(\mathbb{C}); \quad \|A\|_M = 0 \;\Leftrightarrow\; A = 0;$

(2) $\|\alpha A\|_M = |\alpha| \|A\|_M, \quad \forall A \in \mathbb{M}^n(\mathbb{C}), \;\; \forall \alpha \in \mathbb{C};$

(3) $\|A + B\|_M \leq \|A\|_M + \|B\|_M, \quad \forall A, B \in \mathbb{M}^n(\mathbb{C});$

(4) $\|Ax\|_V \leq \|A\|_M \|x\|_V, \quad \forall A \in \mathbb{M}^n(\mathbb{C}), \;\; \forall x \in \mathbb{C}^n;$

(5) $\|AB\|_M \leq \|A\|_M \|B\|_M, \quad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$

Dem.: $(\cdots)$

Proposição. Sendo M a norma associada à norma vectorial V e I a matriz identidade, então $\|I\|_M = 1$.

Exemplo. Normas matriciais em $\mathbb{M}^n(\mathbb{C})$ associadas a normas vectoriais em $\mathbb{C}^n$:

$x = [x_i] \in \mathbb{C}^n$	$A = [a_{ij}] \in \mathbb{M}^n(\mathbb{C})$
$\|x\|_1 = \sum_{i=1}^{n} \|x_i\|$	$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \|a_{ij}\|$ **norma por colunas**
$\|x\|_2 = \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$	$\|A\|_2 = \sqrt{r_\sigma(A^*A)}$ **norma Euclideana**
$\|x\|_\infty = \max_{1\leq i\leq n} \|x_i\|$	$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \|a_{ij}\|$ **norma por linhas**

Nota. A^* designa a matriz transposta conjugada da matriz A.

Definição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ e seja $\sigma(A)$ o **espectro** de A, isto é, o conjunto dos valores próprios de A. Chama-se **raio espectral** de A e representa-se por $r_\sigma(A)$ a

$$r_\sigma(A) = \max_{\lambda\in\sigma(A)} |\lambda|.$$

Proposição. A norma matricial associada à norma da soma em $\mathbb{C}^n$ tem a forma

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|.$$

Dem.: $(\cdots)$

Proposição. A norma matricial associada à norma do máximo em $\mathbb{C}^n$ tem a forma

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Dem.: (Exercício)

Proposição. A norma matricial associada à norma Euclidiana em $\mathbb{C}^n$ tem a forma

$$\|A\|_2 = \sqrt{r_\sigma(A^*A)}.$$

Nota. O cálculo de $\|A\|_2$ é muito mais complicado do que o de $\|A\|_1$ ou $\|A\|_\infty$. Se apenas precisarmos de uma estimativa de $\|A\|_2$ podemos recorrer às desigualdades seguintes.

Proposição. Sendo $A \in \mathbb{M}^n(\mathbb{C})$, então:

(1) $\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\, \|A\|_1$

(2) $\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\, \|A\|_\infty$

(3) $\|A\|_2 \leq \sqrt{\|A\|_1 \, \|A\|_\infty}$

(4) $\frac{1}{\sqrt{n}} \|A\|_F \leq \|A\|_2 \leq \|A\|_F$.

Dem.: (Exercício)

Nota. Sendo $A \in \mathbb{M}^n(\mathbb{C})$, $A = [a_{ij}]$, define-se a **norma de Frobenius** de A por

$$\|A\|_F = \left[\sum_{i,j=1}^{n} |a_{ij}|^2 \right]^{1/2}.$$

Esta norma é regular, isto é,

$$\|AB\|_F \leq \|A\|_F \|B\|_F, \qquad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$$

Esta norma é compatível com a norma euclideana para vectores em $\mathbb{C}^n$, isto é,

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2, \qquad \forall A \in \mathbb{M}^n(\mathbb{C}), \ \ \forall x \in \mathbb{C}^n.$$

Esta norma <u>não</u> está associada a nenhuma norma vectorial pois $\|I\|_F = \sqrt{n}$.

Exemplo. Sendo

$$A = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & -2 & 2 \end{bmatrix}$$

calcular $\|A\|_1, \|A\|_\infty, \|A\|_F, r_\sigma(A), \|A\|_2$.

Resolução:

$\|A\|_1 = \max\{4, 6, 4\} = 6$

$\|A\|_\infty = \max\{4, 5, 5\} = 5$

$\|A\|_F = (4 + 1 + 1 + 1 + 9 + 1 + 1 + 4 + 4)^{1/2} = \sqrt{26} = 5.09902$

$r_\sigma(A) = \max\limits_{\lambda \in \sigma(A)} |\lambda|$

$$\begin{aligned}\sigma(A) &= \{\lambda \in \mathbb{C} : \ -\lambda^3 + 7\lambda^2 - 18\lambda + 18 = 0\} \\ &= \{3, \ 2 + i\sqrt{2}, \ 2 - i\sqrt{2}\}\end{aligned}$$

$$r_\sigma(A) = 3$$

$$\|A\|_2 = \sqrt{r_\sigma(A^T A)}$$

$$\begin{aligned}\sigma(A^T A) &= \{\lambda \in \mathbb{C} : \ -\lambda^3 + 26\lambda^2 - 186\lambda + 324 = 0\} \\ &= \{2.58018, \ 8.31148, \ 15.1083\}\end{aligned}$$

$$\|A\|_2 = 3.88695$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$. Então:

(1) Qualquer que seja a norma matricial M associada a uma certa norma vectorial em $\mathbb{C}^n$, verifica-se
$$r_\sigma(A) \leq \|A\|_M.$$

(2) Qualquer que seja $\varepsilon > 0$ existe uma norma $N(\varepsilon)$ associada a uma certa norma vectorial em $\mathbb{C}^n$ tal que
$$\|A\|_{N(\varepsilon)} \leq r_\sigma(A) + \varepsilon.$$

Dem.: (1) $(\cdots)$

Corolário. Seja $A \in \mathbb{M}^n(\mathbb{C})$. Então $r_\sigma(A) < 1$ se e só se $\|A\|_M < 1$ para alguma norma matricial M associada a uma norma vectorial em $\mathbb{C}^n$.

Dem.: $(\cdots)$

Nota. O raio espectral de A é pois o ínfimo de todas as normas matriciais de A associadas a normas vectoriais em $\mathbb{C}^n$.

Condicionamento de sistemas lineares

• Consideremos o problema de determinar a solução de um sistema linear

$$Ax = b,$$

onde $A \in \mathbb{M}^n(\mathbb{C}), \det A \neq 0, b \in \mathbb{C}^n$. Os dados do problema são neste caso a matriz A e o vector b. A solução é o vector $x = A^{-1}b$. Pretendemos analisar a forma como os erros nos dados afectam os erros na solução. Para isso consideramos um outro sistema linear

$$\tilde{A}\tilde{x} = \tilde{b},$$

e vamos procurar exprimir o erro relativo da solução do segundo sistema em relação à do primeiro, $\dfrac{x - \tilde{x}}{\|x\|}$, em termos dos erros relativos dos dados do segundo sistema em relação aos do primeiro, isto é, $\dfrac{A - \tilde{A}}{\|A\|}$ e $\dfrac{b - \tilde{b}}{\|b\|}$.

Proposição. Consideremos os sistemas lineares

$$Ax = b, \qquad A\tilde{x} = \tilde{b},$$

onde $b, \tilde{b} \in \mathbb{C}^n$, $A \in \mathbb{M}^n(\mathbb{C})$, $\det A \neq 0$. Então

$$\frac{\|\delta_{\tilde{b}}\|_V}{\mathrm{cond}_M(A)} \leq \|\delta_{\tilde{x}}\|_V \leq \mathrm{cond}_M(A)\, \|\delta_{\tilde{b}}\|_V$$

onde

$$\delta_{\tilde{x}} = \frac{x - \tilde{x}}{\|x\|_V}, \qquad \delta_{\tilde{b}} = \frac{b - \tilde{b}}{\|b\|_V}, \qquad \mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

M designa a norma matricial associada à norma vectorial V.

Dem.: $(\cdots)$

Proposição. Consideremos os sistemas lineares

$$Ax = b, \qquad \tilde{A}\tilde{x} = \tilde{b},$$

onde $b, \tilde{b} \in \mathbb{C}^n$, $A, \tilde{A} \in \mathbb{M}^n(\mathbb{C})$, $\det A \neq 0$ e $\tilde{A}$ é tal que

$$\|A - \tilde{A}\|_M \|A^{-1}\|_M < 1.$$

Então $\tilde{A}$ é não singular e verifica-se a desigualdade

$$\|\delta_{\tilde{x}}\|_V \leq \frac{\mathrm{cond}_M(A)}{1 - \|\delta_{\tilde{A}}\|_M \,\mathrm{cond}_M(A)} \left(\|\delta_{\tilde{A}}\|_M + \|\delta_{\tilde{b}}\|_V\right),$$

onde

$$\delta_{\tilde{x}} = \frac{x - \tilde{x}}{\|x\|_V}, \qquad \delta_{\tilde{b}} = \frac{b - \tilde{b}}{\|b\|_V}, \qquad \delta_{\tilde{A}} = \frac{A - \tilde{A}}{\|A\|_M}, \qquad \mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

M designa a norma matricial associada à norma vectorial V.

Definição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz não singular. Chama-se **número de condição de A relativamente à norma** M a

$$\mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

Chama-se **número de condição de A relativamente ao raio espectral** a

$$\mathrm{cond}_*(A) = r_\sigma(A) r_\sigma(A^{-1}).$$

Nota. O número de condição de A relativamente à norma $p \geq 1$ é designado por $\mathrm{cond}_p(A)$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz não singular. Então:

(1) $\text{cond}_*(A) = \dfrac{\max\limits_{\lambda\in\sigma(A)} |\lambda|}{\min\limits_{\lambda\in\sigma(A)} |\lambda|}$;

(2) $\text{cond}_p(A) \geq \text{cond}_*(A) \geq 1, \quad \forall p \geq 1$;

(3) $A = A^* \;\Rightarrow\; \text{cond}_2(A) = \text{cond}_*(A)$;

(4) $AA^* = I \;\Rightarrow\; \text{cond}_2(A) = \text{cond}_*(A) = 1.$

Dem.: $(\cdots)$

Nota. (1) Uma matriz $A \in \mathbb{M}^n(\mathbb{C})$ tal que $A^* = A$ diz-se uma matriz **hermitiana**. Uma matriz hermitiana real é uma matriz **simétrica.**

(2) Uma matriz $A \in \mathbb{M}^n(\mathbb{C})$ tal que $A^*A = AA^* = I$ diz-se uma matriz **unitária**. Uma matriz unitária real é uma matriz **ortogonal.**

Definição. Um sistema linear com matriz A diz-se **bem condicionado** se $\text{cond}_p(A) \lessapprox 1$ e **mal condicionado** se $\text{cond}_p(A) \gg 1$.

Exemplo. Sendo

$$A = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & -2 & 2 \end{bmatrix}$$

calcular $\text{cond}_1(A), \text{cond}_2(A), \text{cond}_\infty(A), \text{cond}_*(A)$.

Resolução:

$$A^{-1} = \frac{1}{18}\begin{bmatrix} 8 & -4 & -2 \\ 3 & 3 & -3 \\ -1 & 5 & 7 \end{bmatrix}$$

$$\|A^{-1}\|_1 = \frac{1}{18}\max\{12, 12, 12\} = \frac{2}{3}$$

$$\|A^{-1}\|_\infty = \frac{1}{18}\max\{14, 9, 13\} = \frac{7}{9}$$

$$r_\sigma(A^{-1}) = \frac{1}{\min_{\lambda\in\sigma(A)}|\lambda|} = \frac{\sqrt{6}}{6}$$

$$r_\sigma\left((A^{-1})^T A^{-1}\right) = \frac{1}{\min_{\lambda\in\sigma(A^TA)}|\lambda|} = 0.387570$$

$$\|A^{-1}\|_2 = 0.622551$$

$$\text{cond}_1(A) = 6 \times \frac{2}{3} = 4$$

$$\text{cond}_2(A) = 3.88695 \times 0.622551 = 2.41982$$

$$\text{cond}_\infty(A) = 5 \times \frac{7}{9} = 3.88889$$

$$\text{cond}_*(A) = 3 \times \frac{\sqrt{6}}{6} = 1.22474$$

Exemplo. Matriz de Hilbert,

$$H_n \in \mathbb{M}^n(\mathbb{R}), \qquad (H_n)_{ij} = \frac{1}{i+j-1}.$$

◇ A matriz inversa é conhecida explicitamente:

$$\left(H_n^{-1}\right)_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!}$$

◇ A matriz aparece, por exemplo, na aproximação mínimos quadrados de uma função.

◇ A matriz é muito mal condicionada:

n	$\text{cond}_*(H_n)$
2	1.93×10
3	5.24×10^2
4	1.55×10^4
5	4.77×10^5
6	1.50×10^7
7	4.75×10^8
8	1.53×10^{10}
9	4.93×10^{11}
10	1.60×10^{13}

◇ A matriz é utilizada para testar programas para resolver sistemas lineares mal condicionados.

Exemplo. Considerem-se os sistemas

$$H_n x = b_n, \qquad n = 2, 3, 4, 5 \tag{1}$$

onde H_n é a matriz de Hilbert de ordem n e $b_n \in \mathbb{C}^n$ é um vector cujas componentes são todas iguais a 1. Considerem-se também os sistemas

$$\tilde{H}_n \tilde{x} = b_n, \qquad n = 2, 3, 4, 5, \tag{2}$$

onde $\tilde{H}_n$ é a matriz que representa a matrix de Hilbert H_n num sistema de ponto flutuante com 4 dígitos na mantissa. Determinar os erros relativos das soluções dos sistemas (2) em relação aos sistemas (1), com o mesmo valor de n, e interpretar os resultados à luz da desigualdade do teorema anterior.

n	$\|\delta_{\tilde{H}_n}\|_\infty$	$\|\delta_{\tilde{x}}\|_\infty$	$\text{cond}_\infty(H_n)$	$z_n = \|\delta_{\tilde{H}_n}\|_\infty \times \text{cond}_\infty(H_n)$	$\frac{z_n}{1-z_n}$
2	0.222×10^{-4}	0.400×10^{-3}	27	0.600×10^{-3}	0.600×10^{-3}
3	0.182×10^{-4}	0.698×10^{-2}	748	0.136×10^{-1}	0.138×10^{-1}
4	0.366×10^{-4}	0.269	28375	0.104×10^{1}	
5	0.480×10^{-4}	0.914	943656	0.453×10^{2}	

Resolução numérica de sistemas lineares

• *Métodos directos*: a solução exacta (na ausência de erros de arredondamento) é obtida num número finito de passos. São exemplos: o *método de eliminação de Gauss*, de que trataremos brevemente para introduzir a *pesquisa parcial de pivot*; os *métodos de factorização*, a que apenas faremos uma breve referência.

• *Métodos iterativos*: a solução aproximada converge para a solução exacta quando o número de iteradas tende para infinito. São exemplos, de que trataremos seguidamente: *método de Jacobi, método de Gauss-Seidel, método de Jacobi modificado ou com relaxação, método de Gauss-Seidel modificado ou com relaxação ou método SOR.*

Método de eliminação de Gauss com pesquisa parcial de pivot

Definição (pesquisa parcial de pivot). No passo k da eliminação de Gauss considere-se

$$s = \min \left\{ r : \left| a_{rk}^{(k)} \right| = \max_{k \le i \le n} \left| a_{ik}^{(k)} \right| \right\}.$$

s é pois o menor dos índices das linhas $r \geq k$ para os quais é atingido o valor máximo dos valores absolutos dos elementos $\left| a_{ik}^{(k)} \right|$ da coluna k localizados abaixo e sobre a diagonal principal. Se $s > k$ trocam-se as linhas s e k da matriz $A^{(k)}$ e do vector $b^{(k)}$; e prossegue-se com o passo k da eliminação de Gauss. Esta estratégia implica que

$$|m_{ik}| \leq 1, \qquad i = k+1, \ldots, n.$$

(0) $A^{(1)} = A, \quad b^{(1)} = b$

$$\left[\begin{array}{l} i = 1, 2, \ldots, n \\ \qquad \left[\begin{array}{l} j = 1, 2, \ldots, n \\ \qquad a_{ij}^{(1)} = a_{ij} \end{array}\right. \\ \quad b_i^{(1)} = b_i \end{array}\right.$$

(1) Redução da matriz A à forma triangular superior

$A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)}$,

(2) Transformação do 2^{o} membro do sitema

$b^{(1)} \to b^{(2)} \to \cdots \to b^{(n)}$

$$\left[\begin{array}{l} k = 1, 2, \ldots, n-1 \qquad \left(A^{(k)} \to A^{(k+1)}, \quad b^{(k)} \to b^{(k+1)}\right) \\ \qquad \left[\begin{array}{l} s > k, \quad s = \min\left\{r : \left|a_{rk}^{(k)}\right| = \max\limits_{k \le i \le n} \left|a_{ik}^{(k)}\right|\right\}. \\ \qquad \left[\begin{array}{l} j = k, k+1, \ldots, n \\ \qquad a_{kj}^{(k)} \leftrightarrow a_{sj}^{(k)} \end{array}\right. \\ \quad b_k^{(k)} \leftrightarrow b_s^{(k)} \end{array}\right. \\ \quad \left[\begin{array}{l} i = k+1, k+2, \ldots, n \\ \qquad m_{ik} = \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \\ \qquad \left[\begin{array}{l} j = k+1, k+2, \ldots, n \\ \qquad a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \end{array}\right. \\ \quad b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \end{array}\right. \end{array}\right.$$

(3) Resolução do sistema transformado : $\quad A^{(n)} x = b^{(n)}$

$$\left[\begin{array}{l} x_n = \dfrac{b_n^{(n)}}{a_{nn}^{(n)}} \\ \quad \left[\begin{array}{l} i = n-1, n-2, \ldots, 1 \\ \qquad x_i = \dfrac{1}{a_{ii}^{(i)}} \left(b_i^{(i)} - \displaystyle\sum_{j=i+1}^{n} a_{ij}^{(i)} x_j\right) \end{array}\right. \end{array}\right.$$

Nota. Utilização do método de eliminação de Gauss para calcular o determinante de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$, não singular:

$$\det A = (-1)^\nu \det A^{(n)} = (-1)^\nu a_{11}^{(n)} a_{22}^{(n)} \cdots a_{nn}^{(n)},$$

onde ν designa o número de troca de linhas efectuado. Comparemos, do ponto de vista da eficiência computacional, medida pelo número de multiplicações e divisões necessárias para obter o resultado, este método com o método baseado na utilização directa da definição:

$$\det A = \sum_{p \in \Pi} \operatorname{sgn}(p) a_{1\alpha} a_{2\beta} \ldots a_{n\omega},$$

onde Π designa o conjunto de todas as permutações de $(1, 2, \ldots, n)$, $p = (\alpha, \beta, \ldots, \omega)$ e $\operatorname{sgn}(p) = \pm 1$.

n	$N_{\text{def}}(n) = n!(n-1)$	$N_{EG}(n) = \frac{1}{3}(n-1)(n^2+n+3)$	$\tilde{N}_{EG}(n) = \frac{n^3}{3}$
2	2	3	
10	3.27×10^7	339	333
100	9.24×10^{159}	333399	333333

$N_{\text{def}}(n)$ designa o número de multiplicações necessárias para calcular $\det A$ por definição, $N_{EG}(n)$ designa o número de multiplicações e divisões necessárias para obter $\det A$ usando o método de eliminação de Gauss para obter a matriz $A^{(n)}$ e $\tilde{N}_{EG}(n)$ designa o valor aproximado (assimptótico) de $N_{EG}(n)$ para valores elevados de n.

Nota. Utilização do método de eliminação de Gauss para calcular a matriz inversa A^{-1} de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$, não singular. Sendo

$$AA^{-1} = I,$$

e designando por $c_1, \ldots, c_n$ as colunas de A^{-1} e por $e_1, \ldots, e_n$ as colunas da matriz identidade, podemos escrever

$$A[c_1 \;\; c_2 \;\; \cdots \;\; c_n] = [e_1 \;\; e_2 \;\; \cdots \;\; e_n].$$

A matriz A^{-1} pode pois ser obtida determinando as soluções $c_1, \ldots, c_n$ dos n sistemas $Ac_1 = e_1, \ldots, Ac_n = e_n$ por eliminação de Gauss. Note-se que todos estes sistemas têm a mesma matriz A pelo que a parte de redução da matriz à forma triangular superior é comum a todos eles.

Nota. Utilização do método de eliminação de Gauss para obter a **factorização triangular** de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$. Sendo A uma matriz não singular então existe uma matriz de permutação P tal que a matriz PA admite uma única factorização triangular $PA = LU$, onde L é uma matriz triangular inferior com diagonal principal unitária e U é uma matriz triangular superior com todos os elementos na diagonal principal diferentes de zero. Os elementos de L por baixo da diagonal principal na coluna k são os

multiplicadores utilizados no passo k da eliminação de Gauss, isto é,

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ m_{21} & 1 & 0 & \cdots & 0 & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{n,n-1} & 1 \end{bmatrix}$$

A matriz U é a matriz do sistema após a eliminação de Gauss, isto é,

$$U = A^{(n)}.$$

A factorização triangular de uma matriz permite reduzir a resolução de um qualquer sistema linear $Ax = b$ à resolução de dois sistemas lineares com matrizes triangulares, a qual se faz facilmente:

$$Ax = b \;\Leftrightarrow\; LUx = Pb \;\Leftrightarrow\; \begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

Exemplo. Considere-se o sistema linear

$$\begin{bmatrix} 1 & 5 \\ 500 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}.$$

(a) Determinar a solução exacta do sistema pelo método de eliminação de Gauss.

(b) Supondo que se efectuam os cálculos num sistema FP(10, 3, -10, 10), com arredondamento simétrico, determinar a solução aproximada do sistema pelo método de eliminação de Gauss.

(c) Idem, pelo método de eliminação de Gauss com pesquisa parcial de pivot.

(d) Comparar os erros relativos das soluções obtidas nas alíneas anteriores.

Resolução:

(a)

$$\begin{bmatrix} 1 & 5 \\ 500 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 5 \\ 0 & -2499 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ -2498 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \dfrac{5}{2499} \\ \dfrac{2498}{2499} \end{bmatrix} = \begin{bmatrix} 0.200080032\cdots \times 10^{-2} \\ 0.999599840\cdots \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1.00 & 5.00 \\ 500. & 1.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 5.00 \\ 2.00 \end{bmatrix}$$

$$\begin{bmatrix} 1.00 & 5.00 \\ 0.00 & -2500. \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 5.00 \\ -2500. \end{bmatrix}$$

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 0.00 \\ 1.00 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 500. & 1.00 \\ 1.00 & 5.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 2.00 \\ 5.00 \end{bmatrix}$$

$$\begin{bmatrix} 500. & 1.00 \\ 0.00 & 5.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 2.00 \\ 5.00 \end{bmatrix}$$

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 0.00200 \\ 1.00 \end{bmatrix}$$

(d) (b) $|\delta_{\tilde{x}}| = 1.$, $|\delta_{\tilde{y}}| = 0.0004$

(c) $|\delta_{\tilde{x}}| = 0.0004$, $|\delta_{\tilde{y}}| = 0.0004$

Métodos iterativos

• Vamos considerar métodos iterativos para obter valores aproximados da solução do sistema linear $Ax = b$, $A \in \mathbb{M}^n(\mathbb{C}), b \in \mathbb{C}^n$, da forma

$$\begin{cases} x^{(k+1)} = Cx^{(k)} + w, & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n. \end{cases}$$

Tratam-se pois de métodos iterativos a um passo com função iteradora $\Phi : \mathbb{C}^n \to \mathbb{C}^n$ definida por $\Phi(x) = Cx + w$, onde $C \in \mathbb{M}^n(\mathbb{C})$ e $w \in \mathbb{C}^n$. C é a **matriz iteradora** do método.

Definição. O método iterativo com função iteradora Φ diz-se **consistente** com o sistema $Ax = b$ se este sistema e o sistema $x = \Phi(x)$ tiverem a mesma solução.

Proposição. O método iterativo com função iteradora Φ é consistente com o sistema $Ax = b$ se e só se

$$(I - C)A^{-1}b = w.$$

Proposição. O método iterativo com função iteradora Φ tal que

$$C = -M^{-1}N = I - M^{-1}A, \qquad w = M^{-1}b,$$

onde $M, N \in \mathbb{M}^n(\mathbb{C})$, M invertível, $M + N = A$, é consistente.

Dem.:

$$Ax = b \;\Rightarrow\; (M + N)x = b \;\Rightarrow\; Mx = -Nx + b \;\Rightarrow\; x = -M^{-1}Nx + M^{-1}b$$

Nota. A forma "intermédia" do método iterativo

$$Mx^{(k+1)} = -Nx^{(k)} + b, \qquad k \in \mathbb{N},$$

será útil mais adiante.

Proposição. Qualquer matriz $A \in \mathbb{M}^n(\mathbb{C})$, $A = [a_{ij}]$, pode ser decomposta na soma

$$A = L_A + D_A + U_A,$$

onde

(i) $L_A = [l_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz estritamente triangular inferior com elementos

$$l_{ij} = a_{ij}, \quad i > j, \qquad l_{ij} = 0, \quad i \leq j.$$

(ii) $D_A = [d_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz diagonal com elementos

$$d_{ij} = a_{ij}, \quad i = j, \qquad d_{ij} = 0, \quad i \neq j.$$

(iii) $U_A = [u_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz estritamente triangular superior com elementos

$$u_{ij} = a_{ij}, \quad i < j, \qquad u_{ij} = 0, \quad i \geq j.$$

Definição. O **método de Jacobi** corresponde a tomar

$$M = D, \qquad N = L + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que D seja invertível. Toma pois a forma:

$$\begin{cases} x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = \dfrac{1}{a_{ii}} \left[b_i - \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], & i = 1, \ldots, n, \qquad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, & i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_J = -D^{-1}(L+U) = I - D^{-1}A.$$

Nota. O método de Jacobi é também conhecido por *método das substituições simultâneas.*

Definição. O **método de Gauss-Seidel** corresponde a tomar

$$M = D + L, \qquad N = U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que $D + L$ seja invertível. Toma pois a forma;

$$\begin{cases} x^{(k+1)} = D^{-1}\left[b - Lx^{(k+1)} - Ux^{(k)}\right], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left[b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=1+1}^{n} a_{ij}x_j^{(k)}\right], & i = 1, \ldots, n, \qquad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \qquad i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_{GS} = -(D + L)^{-1}U = I - (D + L)^{-1}A.$$

Nota. O método de Gauss-Seidel é também conhecido por *método das substituições sucessivas.*

Definição. O **método de Jacobi modificado** ou **método de Jacobi com relaxação** corresponde a tomar

$$M = \frac{D}{\omega}, \qquad N = \left(1 - \frac{1}{\omega}\right) D + L + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que D seja invertível. ω é um parâmetro real positivo conhecido por **parâmetro de relaxação**. O método toma pois a forma:

$$\begin{cases} x^{(k+1)} = (1 - \omega)x^{(k)} + \omega D^{-1}[b - (L + U)x^{(k)}], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \dfrac{\omega}{a_{ii}}\left[b_i - \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(k)}\right], & i = 1, \ldots, n, \qquad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \qquad i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_{Jm}(\omega) = I - \omega D^{-1}A.$$

Nota. Para $\omega = 1$ o método reduz-se ao método de Jacobi.

Definição. O **método de Gauss-Seidel modificado** ou ou **método de Gauss-Seidel com relaxação** ou **método SOR** corresponde a tomar

$$M = \frac{D}{\omega} + L, \qquad N = \left(1 - \frac{1}{\omega}\right) D + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que $D/\omega + L$ seja invertível. ω é um parâmetro real positivo conhecido por **parâmetro de relaxação**. O método SOR toma pois a forma;

$$\begin{cases} x^{(k+1)} = (1-\omega)x^{(k)} + \omega D^{-1}\left[b - Lx^{(k+1)} - Ux^{(k)}\right], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \dfrac{\omega}{a_{ii}}\left[b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=1+1}^{n} a_{ij}x_j^{(k)}\right], \quad i = 1,\ldots,n, \quad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \qquad i = 1,\ldots,n. \end{cases}$$

A matriz iteradora é

$$C_{SOR}(\omega) = I - \omega(D + \omega L)^{-1}A.$$

Nota. SOR é o acrónimo de *successive over-relaxation*. Para $\omega = 1$ o método SOR reduz-se ao método de Gauss-Seidel.

Nota. A introdução do parâmetro de relaxação poderá permitir tornar convergente um método divergente ou acelerar a convergência de um método convergente.

Convergência dos métodos iterativos

• Consideremos o método iterativo consistente com o sistema linear $Ax = b$, $A = M+N$,

$$\begin{cases} x^{(k+1)} = Cx^{(k)} + w, & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases} \qquad (\#)$$

onde

$$C = -M^{-1}N = I - M^{-1}A, \qquad w = M^{-1}b.$$

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, isto é,

$$\lim_{k\to\infty} x^{(k)} = x, \qquad \forall x^{(0)} \in \mathbb{C}^n,$$

ou

$$\lim_{k\to\infty} e^{(k)} = 0, \qquad \forall x^{(0)} \in \mathbb{C}^n,$$

onde $e^{(k)} = x - x^{(k)}$, se e só se

$$\lim_{k\to\infty} C^k e^{(0)} = 0, \qquad \forall e^{(0)} \in \mathbb{C}^n.$$

Dem.: De $x = \Phi(x)$ e $x^{(k+1)} = \Phi(x^{(k)})$ obtém-se por subtracção $e^{(k+1)} = Ce^{(k)}$ e por indução $e^{(k)} = C^k e^{(0)}$.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se existir uma norma matricial M, associada a uma certa norma vectorial V em $\mathbb{C}^n$, tal que $\|C\|_M < 1$. Neste caso são válidas as seguintes estimativas de erro:

(1) $\|e^{(k+1)}\|_V \leq c\|e^{(k)}\|_V$;

(2) $\|e^{(k)}\|_V \leq c^k\|e^{(0)}\|_V$;

(3) $\|e^{(k)}\|_V \leq \dfrac{1}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V$;

(4) $\|e^{(k+1)}\|_V \leq \dfrac{c}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V$;

(5) $\|e^{(k)}\|_V \leq \dfrac{c^k}{1-c}\,\|x^{(1)} - x^{(0)}\|_V$;

onde $e^{(k)} = x - x^{(k)},\ k \in \mathbb{N}$, e $c = \|C\|_M$.

Dem.: $(\cdots)$

Nota. A estimativa "a posteriori"(4) pode ser usada como critério de paragem do método iterativo:

$$\|e^{(k+1)}\|_V \leq \frac{c}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V < \varepsilon$$

$$\|x^{(k+1)} - x^{(k)}\|_V < \varepsilon\left(\frac{1}{c} - 1\right).$$

Repare-se que

$$\begin{cases} 1 \leq \dfrac{1}{c} - 1, & 0 < c \leq \dfrac{1}{2}, \\ 0 < \dfrac{1}{c} - 1 \leq 1, & \dfrac{1}{2} \leq c < 1. \end{cases}$$

Nota. Todas estas estimativas permanecem válidas se substituirmos c por uma outra quantidade $\tilde{c}$ tal que $c < \tilde{c} < 1$. Em particular, $\|C\|_2$, que é difícil de calcular, pode ser substituída por $\|C\|_F < 1$. Recorde-se que $\|C\|_2 \leq \|C\|_F$, $\forall C \in \mathbb{M}^n(\mathbb{C})$.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se $r_\sigma(C) < 1$.

Dem.: $(\cdots)$

Nota. Estes dois teoremas, conjugados com um teorema anterior que nos diz que o raio espectral de uma matriz é o ínfimo de todas as normas associadas a normas vectoriais da matriz, sugerem que o raio espectral determina a rapidez de convergência do método, sendo a convergência tanto mais rápida quanto menor for o raio espectral da matriz iteradora. Mais precisamente pode mostrar-se que:

Proposição. Para o método iterativo (#) os erros $e^{(k)} = x - x^{(k)}$ satisfazem a

$$\sup_{e^{(0)} \in \mathbb{C}^n \setminus \{0\}} \limsup_{k \to \infty} \left(\frac{\|e^{(k)}\|_V}{\|e^{(0)}\|_V} \right)^{1/k} = r_\sigma(C),$$

onde V designa qualquer norma em $\mathbb{C}^n$.

Nota.

(a) Este resultado significa que para valores de k elevados, podemos escrever em muitos casos

$$\|e^{(k+1)}\| \approx r_\sigma(C) \|e^{(k)}\|.$$

(b) Partindo de $x^{(k+1)} = Cx^{(k)} + w$ e de $x^{(k)} = Cx^{(k-1)} + w$ obtém-se

$$x^{(k+1)} - x^{(k)} = C\left(x^{(k)} - x^{(k-1)}\right),$$

isto é, a diferença de duas iteradas consecutivas satisfaz à mesma igualdade que os erros de duas iteradas consecutivas:

$$e^{(k+1)} = Ce^{(k)}.$$

Para valores de k elevados verifica-se também

$$\|x^{(k+1)} - x^{(k)}\| \approx r_\sigma(C) \|x^{(k)} - x^{(k-1)}\|.$$

(c) Na prática, para obter estimativas de erro, utiliza-se em vez de $c = \|C\|$ e de $r_\sigma(C)$ uma outra constante c_r tal que,

$$r^{(k)} = \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V} \leq c_r, \qquad \forall k > k_0,$$

quando for possível obtê-la experimentalmente.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se todas as raízes $\lambda_1, \ldots, \lambda_n$ da equação polinomial

$$\det(\lambda M + N) = 0,$$

tiverem módulo inferior à unidade.

Dem.: $(\cdots)$

Proposição. Os métodos de Jacobi e Gauss-Seidel convergem para a solução do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se $r_\sigma(C_J) < 1$ e $r_\sigma(C_{GS}) < 1$, respectivamente.

Proposição. Os métodos de Jacobi e Gauss-Seidel convergem para a solução do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se todas as raízes das equações

$$\det(\lambda M_J + N_J) = 0, \qquad \det(\lambda M_{GS} + N_{GS}) = 0,$$

respectivamente, tiverem módulo inferior à unidade.

Definição.

(1) Diz-se que a matriz $A \in \mathbb{M}^n(\mathbb{C}), A = [a_{ij}]$ é uma **matriz de diagonal estritamente dominante por linhas** (MDEDL) se verificar as condições

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad \forall i \in \{1, 2, \ldots, n\}.$$

(2) Diz-se que a matriz $A \in \mathbb{M}^n(\mathbb{C}), A = [a_{ij}]$ é uma **matriz de diagonal estritamente dominante por colunas** (MDEDC) se verificar as condições

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|, \qquad \forall j \in \{1, 2, \ldots, n\}.$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL ou MDEDC. Então A é não singular.

Dem.: $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Definam-se as quantidades $\alpha_1, \ldots, \alpha_n$, $\beta_1, \ldots, \beta_n$ por

$$\alpha_1 = 0, \qquad \alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \qquad i \in \{2, \ldots, n\},$$

$$\beta_i = \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|, \qquad i \in \{1, \ldots, n-1\}, \qquad \beta_n = 0.$$

Então:

$$(1)\ \ \alpha_i + \beta_i < 1, \quad \forall i \in \{1, \ldots, n\}; \qquad (2)\ \ \frac{\beta_i}{1 - \alpha_i} < 1, \quad \forall i \in \{1, \ldots, n\}.$$

Dem.: $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Definam-se as quantidades

$$\mu = \max_{1 \leq i \leq n} (\alpha_i + \beta_i) < 1, \qquad \eta = \max_{1 \leq i \leq n} \frac{\beta_i}{1 - \alpha_i} < 1.$$

Então

$$(1)\ \ \|C_J\|_\infty = \mu; \qquad (2)\ \ \eta \leq \mu; \qquad (3)\ \ \|C_{GS}\|_\infty \leq \eta.$$

Dem.: (1) $(\cdots)$ (2) $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Então o método de Jacobi converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, e é válida a estimativa de erro

$$\|e^{(k+1)}\|_\infty \leq \mu \|e^{(k)}\|_\infty.$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Então o método de Gauss-Seidel converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, e é válida a estimativa de erro

$$\|e^{(k+1)}\|_\infty \leq \eta \|e^{(k)}\|_\infty.$$

Nota. Uma vez que $r_\sigma(C_J) \leq \|C_J\|_\infty$ e $r_\sigma(C_{GS}) \leq \|C_{GS}\|_\infty$ este resultado pode levar a pensar que sendo A uma MDEDL então também se verifica que $r_\sigma(C_{GS}) \leq r_\sigma(C_J) < 1$, isto é, que o método de Gauss-Seidel converge pelo menos tão rapidamente quanto o método de Jacobi. Isto não é verdade em geral, mas apenas impondo outras condições a A. Assim, por exemplo, é verdadeiro o seguinte resultado:

Proposição (Teorema de Stein-Rosenberg). Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz tal que a matriz iteradora do método de Jacobi é não negativa, $C_J \geq 0$ (isto é, $(C_J)_{ij} \geq 0, \forall i, j$). Seja C_{GS} a matriz iteradora do método de Gauss-Seidel. Então uma e uma só das seguintes proposições é verdadeira:

(1) $r_\sigma(C_J) = r_\sigma(C_{GS}) = 0$;

(2) $0 < r_\sigma(C_{GS}) < r_\sigma(C_J) < 1$;

(3) $r_\sigma(C_J) = r_\sigma(C_{GS}) = 1$;

(4) $1 < r_\sigma(C_J) < r_\sigma(C_{GS})$.

Nota. A condição $C_J \geq 0$ é satisfeita em particular se A é tal que $D_A > 0$ e $A - D_A \leq 0$. Uma vez que esta condição é verificada para quase todos os sistemas lineares que são obtidos pela aproximação às diferenças finitas de operadores diferenciais lineares, este teorema dá-nos em muitos casos práticos a informação importante de que, quando convergem, o método de Gauss-Seidel converge mais depressa do que o método de Jacobi.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDC. Então quer o método de Jacobi quer o método de Gauss-Seidel convergem para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

Definição. Uma matriz hermitiana $A \in \mathbb{M}^n(\mathbb{C})$ diz-se **definida positiva** se

$$x^* A x > 0, \qquad \forall x \in \mathbb{C}^n \setminus \{0\}.$$

Proposição. Cada uma das seguintes condições é necessária e suficiente para que uma matriz hermitiana $A \in \mathbb{M}^n(\mathbb{C})$ seja definida positiva:

(1) Todos os valores próprios de A são positivos.

(2) Todos os menores principais de A são positivos.

Nota. Chama-se submatriz principal de $A \in \mathbb{M}^n(\mathbb{C})$ à matriz $A_k \in \mathbb{M}^k(\mathbb{C})$, $k \in \{1, 2, \ldots, n\}$ cujos elementos são os elementos das primeiras k linhas e k colunas de A. Chama-se **menor principal** de A a $\det A_k$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva. Então o método de Gauss-Seidel converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva e tal que a matriz $2D_A - A$ é definida positiva. Então o método de Jacobi converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

• A análise de convergência do método SOR é mais complicada que a do método de Jacobi com relaxação pois a matriz iteradora depende não linearmente do parâmetro de relaxação:

$$C_{Jm}(\omega) = I - \omega D^{-1}A, \qquad C_{SOR}(\omega) = I - \omega(D + \omega L)^{-1}A.$$

Proposição. O método de Jacobi modificado é convergente se e só se todos os valores próprios da matriz C_J do método de Jacobi tiverem partes reais inferiores à unidade. A convergência verifica-se para $\omega \in]0, \omega_{\sup}[$, onde

$$\omega_{\sup} = \min_{\lambda \in \sigma(C_J)} \frac{2b(\lambda)}{2b(\lambda) + |\lambda|^2 - 1}, \qquad b(\lambda) = 1 - \Re(\lambda),$$

Proposição. Se o método de Jacobi for convergente então o método de Jacobi modificado com $\omega \in]0, 1]$ também é convergente.

Proposição (Teorema de Kahan). Seja $A \in \mathbb{M}^n(\mathbb{C})$. É condição necessária para que o método SOR convirja para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, que $\omega \in]0, 2[$.

Nota. Prova-se que

$$r_\sigma(C_{SOR}(\omega)) \geq |\omega - 1|, \qquad \forall \omega \in \mathbb{R},$$

e portanto que

$$r_\sigma(C_{SOR}(\omega)) \geq 1, \qquad \forall \omega \notin]0, 2[.$$

Proposição (Teorema de Ostrowski-Reich). Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva. É condição suficiente para que o método SOR convirja para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, que $\omega \in]0, 2[$.

Nota. Prova-se que

$$r_\sigma(C_{SOR}(\omega)) < 1, \qquad \forall \omega \in]0, 2[.$$

Nota. Em particular este resultado significa que o método de Gauss-Seidel, obtido para $\omega = 1$, converge para matrizes hermiteanas e definidas positivas.

• O cálculo do parâmetro de relaxação óptimo, isto é, o parâmetro que minimiza o raio espectral, é difícil excepto nalguns casos particulares. Geralmente obtém-se apenas um valor aproximado experimentando numericamente diversos valores de ω e observando o

efeito na rapidez de convergência. Este esforço é justificado pois o aumento da rapidez de convergência pode ser significativo. Damos seguidamente um exemplo de uma classe de matrizes para a qual o problema pode ser tratado analiticamente com facilidade. Um outro caso que também pode ser tratado analiticamente é o de uma outra classe de matrizes que ocorre na discretização de problemas de valor na fronteira, as chamadas *matrizes consistentemente ordenadas*.

Proposição. Suponhamos que a matriz iteradora C_J do método de Jacobi tem valores próprios reais e inferiores à unidade e designemos por λ_1 e λ_n o maior e o menor destes valores, respectivamente. Então o método de Jacobi modificado com $\omega \in]0, \omega_{\text{sup}}[$, onde $\omega_{\text{sup}} = \dfrac{2}{1-\lambda_n}$, é convergente. Além disso o raio espectral de $C_{Jm}(\omega)$ tem o valor mínimo para $\omega_{\text{opt}} = \dfrac{2}{2-\lambda_1-\lambda_n}$ e esse valor mínimo é

$$r_\sigma(C_{Jm}(\omega_{\text{opt}})) = \frac{\lambda_1 - \lambda_n}{2-\lambda_1-\lambda_n}.$$

Dem.: $(\cdots)$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 10 & 3 & 1 \\ 2 & -10 & 3 \\ 1 & 3 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix}.$$

(a) Determinar as seis primeiras iteradas do método de Jacobi com condição inicial $x^{(0)} = [0 \;\; 0 \;\; 0]^T$. Justificar a convergência do método e obter uma estimativa do erro da iterada $x^{(6)}$.

(b) Idem, para o método de Gauss-Seidel.

Resolução:

(a) $x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}]$

$$x^{(k+1)} = \frac{1}{10}\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix} - \begin{bmatrix} 0 & 3 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{10}\left(14 - 3x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \dfrac{1}{10}\left(5 + 2x_1^{(k)} + 3x_3^{(k)}\right) \\ x_3^{(k+1)} = \dfrac{1}{10}\left(14 - x_1^{(k)} - 3x_2^{(k)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000000	0.000000000	0.000000000
1	1.400000000	0.500000000	1.400000000
2	1.110000000	1.200000000	1.110000000
3	0.929000000	1.055000000	0.929000000
4	0.990600000	0.964500000	0.990600000
5	1.011590000	0.995300000	1.011590000
6	1.000251000	1.005795000	1.000251000

(b) $x^{(k+1)} = D^{-1}[b - Lx^{(k+1)} - Ux^{(k)}]$

$$x^{(k+1)} = \frac{1}{10}\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 3 & 0 \end{bmatrix} x^{(k+1)} - \begin{bmatrix} 0 & 3 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{10}\left(14 - 3x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{10}\left(5 + 2x_1^{(k+1)} + 3x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{10}\left(14 - x_1^{(k+1)} - 3x_2^{(k+1)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000000	0.000000000	0.000000000
1	1.400000000	0.780000000	1.026000000
2	1.063400000	1.020480000	0.987516000
3	0.995104400	0.995275680	1.001906856
4	1.001226610	1.000817379	0.999632125
5	0.999791574	0.999847952	1.000066457
6	1.000038969	1.000027731	0.999987784

Convergência: os métodos de Jacobi e Gauss-Seidel são convergentes para a solução do sistema para qualquer condição inicial pois A é MDEDL.

Estimativas de erro:

$$\|e^{(k)}\|_\infty \leq \frac{c}{1-c}\,\|x^{(k)} - x^{(k-1)}\|_\infty$$

$c = \mu = \max_{1\leq i\leq n}(\alpha_i + \beta_i)$ Método de Jacobi

$c = \eta = \max_{1\leq i\leq n}\dfrac{\beta_i}{1-\alpha_i}$ Método de Gauss-Seidel

i	α_i	β_i	$\alpha_i + \beta_i$	$\dfrac{\beta_i}{1-\alpha_i}$
1	0	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{4}{10}$
2	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{3}{8}$
3	$\frac{4}{10}$	0	$\frac{4}{10}$	0

$$\mu = \frac{1}{2}, \qquad \eta = \frac{2}{5}$$

Método de Jacobi:

$$x^{(6)} - x^{(5)} = \begin{bmatrix} -0.0113390 \\ 0.0104950 \\ -0.0113390 \end{bmatrix}, \qquad \|x^{(6)} - x^{(5)}\|_\infty = 0.0113390$$

$$\|e^{(6)}\|_\infty \leq \|x^{(6)} - x^{(5)}\|_\infty = 0.0113390$$

Com $c = r_\sigma(C_J) = \dfrac{\sqrt{15}}{10} = 0.387298$ obtém-se

$$\|e^{(6)}\|_\infty \leq 0.00716755$$

Estes valores devem ser comparado entre si e com o "erro verdadeiro"

$$\|e^{(6)}\|_\infty = 0.005795$$

Método de Gauss-Seidel:

$$x^{(6)} - x^{(5)} = \begin{bmatrix} 0.000247395 \\ 0.000179779 \\ -0.000078673 \end{bmatrix}, \qquad \|x^{(6)} - x^{(5)}\|_\infty = 0.000247395$$

$$\|e^{(6)}\|_\infty \leq \frac{2}{3}\|x^{(6)} - x^{(5)}\|_\infty = 0.000164930$$

Com $c = r_\sigma(C_{GS}) = \dfrac{\sqrt{67 + \sqrt{13489}}}{1000} = 0.183142$ obtém-se

$$\|e^{(6)}\|_\infty \leq 0.0000554667$$

Estes valores devem ser comparado entre si e com o "erro verdadeiro"

$$\|e^{(6)}\|_\infty = 0.000038969$$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$$

(a) Determinar a solução aproximada $x^{(k)}$ do sistema pelo método de Jacobi com condição inicial $x^{(0)} = [0.5 \;\; 0.8 \;\; 1.0]^T$ tal que é satisfeita a desigualdade

$$\|x^{(k)} - x^{(k-1)}\|_2 \leq 0.01.$$

Obter uma estimativa do erro da iterada $x^{(k)}$.

(b) Idem, para o método de Gauss-Seidel.

Resolução:

(a) $x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}]$

$$x^{(k+1)} = \frac{1}{2}\left(\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(2 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{2}\left(2 + x_1^{(k)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{2}\left(1 + x_2^{(k)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k)} - x^{(k-1)}\|_2$	$\frac{\|x^{(k)} - x^{(k-1)}\|_2}{\|x^{(k-1)} - x^{(k-2)}\|_2}$
0	0.5	0.8	1.0		
1	0.6	0.75	0.9	0.15	
2	0.625	0.85	0.875	0.106066	0.707107
3	0.575	0.875	0.925	0.07500	0.707107
4	0.5625	0.825	0.9375	0.053033	0.707107
5	0.5875	0.8125	0.9125	0.03750	0.707107
6	0.59375	0.8375	0.90625	0.0265165	0.707107
7	0.58125	0.84375	0.91875	0.01875	0.707107
8	0.578125	0.83125	0.921875	0.0132583	0.707107
9	0.584375	0.828125	0.915625	0.009375	0.707107

$$\|e^{(9)}\|_2 \leq \frac{c}{1-c}\|x^{(9)} - x^{(8)}\|_2$$

$$c = r_\sigma(C_J) = \frac{1}{\sqrt{2}}$$

$$\|e^{(9)}\|_2 \leq 0.0242$$

(b) $x^{(k+1)} = D^{-1}[b - Lx^{(k+1)} - Ux^{(k)}]$

$$x^{(k+1)} = \frac{1}{2}\left(\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} x^{(k+1)} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(2 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{2}\left(2 + x_1^{(k+1)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{2}\left(1 + x_2^{(k+1)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k)} - x^{(k-1)}\|_2$	$\frac{\|x^{(k)} - x^{(k-1)}\|_2}{\|x^{(k-1)} - x^{(k-2)}\|_2}$
0	0.5	0.8	1.0		
1	0.6	0.8	0.9	0.141421	
2	0.6	0.85	0.925	0.0559017	0.395285
3	0.575	0.825	0.9125	0.037500	0.67082
4	0.5875	0.8375	0.91875	0.018750	0.5
5	0.58125	0.83125	0.915625	0.009375	0.5

$$\|e^{(5)}\|_2 \leq \frac{c}{1-c}\|x^{(5)} - x^{(4)}\|_2$$

$$c = r_\sigma(C_{GS}) = \frac{1}{2}$$

$$\|e^{(5)}\|_2 \leq 0.01$$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}.$$

(a) Determinar o valor $\omega = \omega_{\text{opt}}$ para o qual o método de Jacobi com relaxação converge mais rapidamente e o valor de $r_\sigma(C_{Jm}(\omega_{\text{opt}}))$.

(b) Idem, para o método de Gauss-Seidel com relaxação.

Resolução:

(a)

$$C_{Jm}(\omega) = I - \omega D^{-1}A = \begin{bmatrix} 1-\omega & -\frac{\omega}{2} & -\frac{\omega}{2} \\ -\frac{\omega}{2} & 1-\omega & -\frac{\omega}{3} \\ -\frac{\omega}{2} & -\omega & 1-\omega \end{bmatrix}$$

$$\sigma\left(C_{Jm}(\omega)\right) = \left\{1 - \frac{\omega}{2}, 1 - \frac{\omega}{2}, 1 - 2\omega\right\}$$

$$r_\sigma\left(C_{Jm}(\omega)\right) = \begin{cases} 1 - 2\omega, & \omega \leq 0, \\ 1 - \frac{\omega}{2}, & 0 \leq \omega \leq \omega_{\text{opt}}, \\ 2\omega - 1, & \omega_{\text{opt}} \leq \omega \end{cases}$$

$\omega_{\text{opt}} = \frac{4}{5}, \qquad r_\sigma\left(C_{Jm}(\omega_{\text{opt}})\right) = \frac{3}{5}$

O método converge para $\omega \in]0,1[$, intervalo em que $r_\sigma\left(C_{Jm}(\omega)\right) < 1$.

(b)

$C_{\text{SOR}}(\omega) = I - \omega(D + \omega L)^{-1}A$

$$= \begin{bmatrix} 1-\omega & -\frac{\omega}{2} & -\frac{\omega}{2} \\ -\frac{\omega}{3}(1-\omega) & \frac{1}{6}(6-6\omega+\omega^2) & -\frac{\omega}{6}(2-\omega) \\ -\frac{\omega}{6}(3-5\omega+2\omega^2) & -\frac{\omega}{12}(12-15\omega+2\omega^2) & \frac{1}{12}(12-12\omega+7\omega^2-2\omega^3) \end{bmatrix}$$

$$\begin{aligned} \det(C_{\text{SOR}}(\omega) - \lambda I) = \; & (1-\omega)^3 + \frac{1}{12}(-36+72\omega-45\omega^2+8\omega^3)\lambda \\ & + \frac{1}{12}(36-36\omega+9\omega^2-2\omega^3)\lambda^2 - \lambda^3 \end{aligned}$$

$r_\sigma(C_{\text{SOR}}(\omega)) = \max_{i=1,2,3}\{|\lambda_i(\omega)\}$

$\omega_{\text{opt}} = \{\omega \in \mathbb{R}^+ : r_\sigma(C_{\text{SOR}}(\omega)) \text{ é mínimo }\}$

ω	$r_\sigma(C_{\text{SOR}}(\omega))$
0.	1.
0.25	0.880321
0.5	0.75
0.75	0.591443
1.	0.333333
1.05	0.252151
1.1	0.239192
1.15	0.252158
1.2	0.279534
1.25	0.313121
1.5	0.5
1.75	0.9061
1.8	0.994896
1.9	1.18037
2.0	1.3767

$\omega_{\text{opt}} = 1.1, \qquad r_\sigma(C_{\text{SOR}}(\omega_{\text{opt}})) = 0.239192$

O método é convergente para $\omega \in]0, 1.8]$.

Determinação experimental do valor óptimo de ω para ambos os métodos:

$$Ax = b, \qquad b = \begin{bmatrix} 4 \\ 5 \\ 5 \end{bmatrix}, \qquad x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

ω	$k < 200 : \|x^{(k)} - x^{(k-1)}\|_2 \leq 10^{-5}$	
	Método Jm	Método SOR
0.1	175	169
0.2	94	86
0.3	64	56
0.4	49	38
0.5	39	26
0.6	33	25
0.7	28	22
0.8	**26**	19
0.9	58	16
1.0		13
1.1		**11**
1.2		12
1.3		14
1.4		17
1.5		19
1.6		31
1.7		64

Comparação entre os métodos directos e os métodos iterativos

Notas.

(a) Os métodos directos requerem $\approx \frac{n^3}{3}$ operações para obter a solução. Os métodos iterativos implicam normalmente um cálculo de $\approx n^2$ operações em cada iteração, o que os torna ineficazes face aos métodos directos para $n_{\text{it}} > \frac{n}{3}$. Por outro lado, a precisão atingida ao fim de $\frac{n}{3}$ iteradas não é normalmente muito boa ($\|x - x^{(n/3)}\|_V \leq c^{n/3}\|x - x^{(0)}\|_V$). Por estas razões os métodos iterativos só se tornam realmente eficazes para matrizes de grandes dimensões e, em especial, quando as matrizes são esparsas (isto é, matrizes com poucos elementos diferentes de zero). Nestes casos os métodos directos não se simplificam em geral muito enquanto que os métodos iterativos apresentam uma redução apreciável do número de operações.

(b) Os métodos iterativos para sistemas lineares convergentes são *estáveis*, isto é, partindo de dois vectores iniciais próximos, $\xi^{(0)}$ e $\eta^{(0)}$, obtêm-se sempre duas sucessões $\{x^{(k)}\}$ e $\{y^{(k)}\}$ igualmente próximas, convergindo ambas para o mesmo vector x, solução exacta do sistema linear, verificando-se:

$$\exists \theta > 0 : \max_k \|x^{(k)} - y^{(k)}\| \leq \theta \|\xi^{(0)} - \eta^{(0)}\|, \quad \forall \xi^{(0)}, \eta^{(0)} \in \mathbb{C}^n.$$

Na presença de erros de arredondamento os métodos iterativos, desde que aplicados a sistemas bem condicionados, continuam estáveis. Isto significa que não há perigo de os erros de arredondamento cometidos no cálculo poderem resultar em erros significativos no resultado final. Isto representa uma importante vantagem dos métodos iterativos sobre os métodos directos em que os erros de arredondamento

podem propagar-se ao longo do cálculo, conduzindo a erros muito grandes no resultado final, mesmo que os sistemas sejam bem condicionados, e sobretudo quando se tratam de sistemas de grandes dimensões.

5. RESOLUÇÃO NUMÉRICA DE SISTEMAS NÃO-LINEARES

• Vamos agora considerar métodos iterativos para resolver sistemas de equações em $\mathbb{R}^n$,

$$f(x) = 0 \quad \Leftrightarrow \quad x = g(x),$$

onde

$$\begin{aligned} x &= [x_1 \;\; x_2 \;\; \ldots \;\; x_n]^T \in \mathbb{R}^n \\ f : D \subset \mathbb{R}^n \to \mathbb{R}^n, &\qquad f(x) = [f_1(x) \;\; f_2(x) \;\; \ldots \;\; f_n(x)]^T, \\ g : D \subset \mathbb{R}^n \to \mathbb{R}^n, &\qquad g(x) = [g_1(x) \;\; g_2(x) \;\; \ldots \;\; g_n(x)]^T. \end{aligned}$$

Definição. Sendo $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$, chama-se **matriz Jacobiana** de f à matriz de derivadas parciais, caso existam,

$$J_f(x) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} (x).$$

Método do ponto fixo

Definição. Diz-se que z é um **ponto fixo** de uma função $g : \mathbb{R}^n \to \mathbb{R}^n$ se e só se $z = g(z)$.

Definição. O **método do ponto fixo** é o método iterativo a um passo da forma

$$x^{(m+1)} = g(x^{(m)}), \quad m \in \mathbb{N}, \qquad x^{(0)} = \xi_0 \in \mathbb{R}^n \;\; \text{(dado)}.$$

Definição. Uma função $g : D \subset \mathbb{R}^n \to \mathbb{R}^n$, diz-se uma **função de Lipschitz** ou **Lipschitziana** em D se existir um número real $L \geq 0$ tal que

$$\|g(x) - g(y)\|_V \leq L\|x - y\|_V, \qquad \forall x, y \in D.$$

onde V designa qualquer norma em $\mathbb{R}^n$. Se $L < 1$ a função diz-se **contractiva** ou uma **contracção** em D, com **constante de contractividade** L.

Teorema do ponto fixo (em $\mathbb{R}^n$). Seja $g : D \to \mathbb{R}^n$ uma função contractiva num conjunto fechado $D \subset \mathbb{R}^n$ com constante de contractividade L e tal que $g(D) \subset D$. Então:

(1) g tem um e um só ponto fixo z em D;

(2) a sucessão $\{x^{(m)}\}_{m\in\mathbb{N}}$, definida por $x^{(m+1)} = g(x^{(m)})$, converge para z, qualquer que seja $x^{(0)} \in D$;

(3) verificam-se as seguintes estimativas de erro:

(i) $\|e^{(m+1)}\|_V \leq L\|e^{(m)}\|_V$;

(ii) $\|e^{(m)}\|_V \leq L^m\|e^{(0)}\|_V$;

(iii) $\|e^{(m)}\|_V \leq \dfrac{1}{1-L}\|x^{(m+1)} - x^{(m)}\|_V$;

(iv) $\|e^{(m+1)}\|_V \leq \dfrac{L}{1-L}\|x^{(m+1)} - x^{(m)}\|_V$;

(v) $\|e^{(m)}\|_V \leq \dfrac{L^m}{1-L}\|x^{(1)} - x^{(0)}\|_V$;

onde $e^{(m)} = z - x^{(m)}$, $\forall m \in \mathbb{N}$ e V designa qualquer norma em $\mathbb{R}^n$.

Dem.: $(\cdots)$

Definição. Um conjunto $D \subset \mathbb{R}^n$ diz-se **convexo** se

$$tx + (1-t)y \in D, \qquad \forall t \in [0,1], \qquad \forall x, y \in D,$$

isto é, se $x, y \in D$ então todos os pontos do segmento de recta que une x e y também pertencem a D.

Proposição. Seja $g \in C^1(D)$ onde D é um conjunto convexo em $\mathbb{R}^n$. Então se

$$\sup_{x \in D} \|J_g(x)\|_M \leq L < 1,$$

a função g é contractiva em D segundo a norma M associada à norma vectorial V em $\mathbb{R}^n$ com constante de contractividade L.

Proposição. Seja D um conjunto fechado e convexo em $\mathbb{R}^n$ e $g \in C^1(D)$ uma função tal que:

(i) $g(D) \subset D$; (ii) $\sup_{x \in D} \|J_g(x)\|_M \leq L < 1$.

Então são válidas as conclusões do teorema do ponto fixo.

Proposição (da convergência local). Seja $g \in C^1(V_z)$, onde V_z é uma vizinhança de z, ponto fixo de g, tal que $r_\sigma(J_g(z)) < 1$. Então o método do ponto fixo converge para z desde que $x^{(0)}$ esteja suficientemente perto de z. Além disso as estimativas de erro (i)-(v) são válidas em alguma bola fechada contida em V_z e que contenha z.

Exemplo. Considere o sistema de equações algébricas não-lineares

$$f(x) = 0, \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad f(x) = \begin{bmatrix} 3x_1 + x_2^2 \\ x_1^2 + 3x_2 - 1 \end{bmatrix}. \tag{S}$$

(a) Mostre que o sistema (S) tem uma e uma só solução z no conjunto

$$D = \left[-\frac{1}{3}, 0\right] \times \left[0, \frac{1}{3}\right].$$

(b) Mostre que o sistema (S) tem uma e uma só solução $\tilde{z}$ no conjunto

$$\tilde{D} = [-4, -2] \times [-4, -2].$$

(c) Utilize o método do ponto fixo para obter um valor aproximado da raiz z com um erro absoluto inferior a 10^{-6}.

(d) Mostre que o sistema (S) tem apenas duas soluções, z e $\tilde{z}$, em $\mathbb{R}^2$.

Resolução:

(a)

$$f(x) = 0 \quad \Leftrightarrow \quad x = g(x), \qquad g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} -\frac{x_2^2}{3} \\ \frac{1}{3}\left(1 - x_1^2\right) \end{bmatrix}$$

O teorema do ponto fixo diz-nos que g terá um e um só ponto fixo em D se forem satisfeitas as seguintes condições:

$$\text{(i) } g \in C^1(D) \qquad \text{(ii) } g(D) \subset D \qquad \text{(iii) } \max_{x \in D} \|J_g(x)\|_\infty < 1$$

Verifiquemos pois estas condições.

(i) g_1 e g_2 são continuamente diferenciáveis em $\mathbb{R}^2$ e portanto em D.

$$\left.\begin{array}{l} \text{(ii) } g_1(x) \in \left[-\frac{1}{27}, 0\right] \subset \left[-\frac{1}{3}, 0\right], \quad \forall x \in D \\ \qquad g_2(x) \in \left[\frac{8}{27}, \frac{1}{3}\right] \subset \left[0, \frac{1}{3}\right], \quad \forall x \in D \end{array}\right\} \quad \Rightarrow \quad g(D) \subset D$$

$$\text{(iii) } J_g(x) = \begin{bmatrix} 0 & -\frac{2x_2}{3} \\ -\frac{2x_1}{3} & 0 \end{bmatrix}$$

$$\max_{x \in D} \|J_g(x)\|_\infty = \max_{x \in D} \max\left\{\frac{2|x_2|}{3}, \frac{2|x_1|}{3}\right\} = \frac{2}{9} < 1$$

(b)

$$f(x) = 0 \quad \Leftrightarrow \quad x = \tilde{g}(x), \qquad \tilde{g}(x) = \begin{bmatrix} \tilde{g}_1(x) \\ \tilde{g}_2(x) \end{bmatrix} = \begin{bmatrix} -\sqrt{1 - 3x_2} \\ -\sqrt{-3x_1} \end{bmatrix}$$

O teorema do ponto fixo diz-nos que $\tilde{g}$ terá um e um só ponto fixo em $\tilde{D}$ se forem satisfeitas as seguintes condições:

$$\text{(i) } \tilde{g} \in C^1(\tilde{D}) \qquad \text{(ii) } \tilde{g}(\tilde{D}) \subset \tilde{D} \qquad \text{(iii) } \max_{x \in \tilde{D}} \|J_{\tilde{g}}(x)\|_\infty < 1$$

Verifiquemos pois estas condições.

(i) $\tilde{g}_1$ e $\tilde{g}_2$ são continuamente diferenciáveis em $]-\infty, 0[\times]-\infty, \frac{1}{3}[$ e portanto em $\tilde{D}$.

(ii) $\left.\begin{array}{ll} \tilde{g}_1(x) \in \left[-\sqrt{13}, -\sqrt{7}\right] \subset [-4,-2], & \forall x \in \tilde{D} \\ \tilde{g}_2(x) \in \left[-\sqrt{12}, -\sqrt{6}\right] \subset [-4,-2], & \forall x \in \tilde{D} \end{array}\right\} \Rightarrow \tilde{g}(\tilde{D}) \subset \tilde{D}$

(iii) $J_{\tilde{g}}(x) = \begin{bmatrix} 0 & \frac{3/2}{\sqrt{1-3x_2}} \\ \frac{3/2}{\sqrt{-3x_1}} & 0 \end{bmatrix}$

$$\max_{x\in\tilde{D}} \|J_{\tilde{g}(x)}\|_\infty = \max_{x\in\tilde{D}} \max\left\{\frac{3/2}{\sqrt{1-3x_2}}, \frac{3/2}{\sqrt{-3x_1}}\right\} = \frac{\sqrt{6}}{4} < 1$$

(c)

$$x^{(m+1)} = g\left(x^{(m)}\right), \quad m \in \mathbb{N}, \qquad x^{(0)} \in D$$

$$\|z - x^{(m)}\|_\infty \leq \frac{L}{1-L}\|x^{(m)} - x^{(m-1)}\|_\infty \equiv B_m$$

$$L = \max_{x\in D} \|J_g(x)\|_\infty = \frac{2}{9}$$

m	$x_1^{(m)}$	$x_2^{(m)}$	$\|x^{(m)} - x^{(m-1)}\|_\infty$	B_m
0	0.000000000000000	0.000000000000000		
1	0.000000000000000	0.333333333333333	0.333×10^{0}	0.952×10^{-1}
2	−0.037037037037037	0.333333333333333	0.370×10^{-1}	0.106×10^{-1}
3	−0.037037037037037	0.332876085962506	0.457×10^{-3}	0.131×10^{-3}
4	−0.036935496201906	0.332876085962506	0.102×10^{-3}	0.290×10^{-4}
5	−0.036935496201906	0.332878589706773	0.250×10^{-5}	0.715×10^{-6}
6	−0.036936051828390	0.332878589706773		
7	−0.036936051828390	0.332878576025110		
8	−0.036936048792168	0.332878576025110		
9	−0.036936048792168	0.332878576099874		
10	−0.036936048808760	0.332878576099874		
11	−0.036936048808760	0.332878576099466		
12	−0.036936048808669	0.332878576099466		
13	−0.036936048808669	0.332878576099468		
14	−0.036936048808670	0.332878576099468		

(d)

$$\left\{\begin{array}{l} 3x_1 + x_2^2 = 0 \\ x_1^2 + 3x_2 - 1 = 0 \end{array}\right. \qquad \left\{\begin{array}{l} x_1 = -\dfrac{x_2^2}{3} \\ x_2^4 + 27x_2 - 9 = 0 \end{array}\right.$$

$$h(t) = t^4 + 27t - 9$$

$$h'(t) = 4t^3 + 27, \qquad h''(t) = 12t^2$$

$h'(t) = 0 \;\Leftrightarrow\; t = -\dfrac{3}{\sqrt[3]{4}} \equiv t_m, \qquad h(t_m) \approx -47.2701$

$h(t) \to \infty$ quando $|t| \to \infty$

$\Rightarrow$ h tem apenas duas raízes em $\mathbb{R}$.

$\Rightarrow$ O sistema dado tem apenas duas raízes em $\mathbb{R}^2$.

Método de Newton generalizado

Definição. O **método de Newton** é o método iterativo a um passo da forma

$$x^{(m+1)} = x^{(m)} - [J_f(x^{(m)})]^{-1} \cdot f(x^{(m)}), \quad m \in \mathbb{N}, \qquad x^{(0)} = \xi_0 \text{ (dado)}.$$

Na prática resolve-se em cada iteração o sistema linear

$$J_f(x^{(m)}) \cdot \Delta x^{(m)} = -f(x^{(m)}),$$

e define-se a iterada seguinte por

$$x^{(m+1)} = x^{(m)} + \Delta x^{(m)}.$$

Teorema de Kantorovich (condições suficientes de convergência).[1]

Seja $D \subset \mathbb{R}^n$ um conjunto aberto e convexo e $f \in C^1(D)$. Suponhamos que para alguma norma em $\mathbb{R}^n$ e algum $x^{(0)} \in D$ são verificadas as seguintes condições:

(i) $\det(J_f(x)) \neq 0, \;\; \forall x \in D; \quad \exists M_1 > 0 : \|[J_f(x)]^{-1}\| \leq \dfrac{1}{M_1}, \;\; \forall x \in D;$

(ii) $\exists M_2 > 0 : \|J_f(x) - J_f(y)\| \leq M_2 \|x - y\|, \;\; \forall x, y \in D$;

(iii) sendo $\varepsilon_0 = 2\|x^{(1)} - x^{(0)}\|$ e $K = \dfrac{M_2}{2M_1}$, verifica-se a desigualdade $2K\varepsilon_0 < 1$;

(iv) $\bar{B}(x^{(0)}, \varepsilon_0) \subset D$.

Então:

(1) f tem um único zero $z \in \bar{B}(x^{(0)}, \varepsilon_0)$;

(2) a sucessão de Newton com condição inicial $x^{(0)}$ é bem definida, permanece em $\bar{B}(x^{(0)}, \varepsilon_0)$ e converge para z.

(3) verifica-se a estimativa de erro a priori

$$\|z - x^{(m)}\| \leq \frac{1}{K} \left(K\varepsilon_0\right)^{2^m}.$$

[1] Este teorema e a sua aplicação ao cálculo de estimativas de erro das iteradas do método de Newton, ilustrada no exemplo que encerra este capítulo, não aparecerão nas provas escritas de avaliação.

Nota. Se $f \in C^2(D)$ a condição (ii) pode ser substituída pela condição:

$$\text{(ii)'} \quad \exists M_2 > 0 : \|H_{f_i}(x)\| \leq M_2, \quad \forall x \in D, \ \ \forall i \in \{1, \ldots, n\},$$

onde f_i designa uma componente de f e H_{f_i} é a matriz Hessiana de f_i, com componentes

$$(H_{f_i})_{jk} = \frac{\partial^2 f_i}{\partial x_j \partial x_k}.$$

Notas.

- ◇ A condição (i) implica a existência de inversa para a matriz Jacobiana (equivalente no caso real a $f'(x) \neq 0$), e serve ao mesmo tempo para definir M_1 (que corresponde no caso real a $\min |f'(x)|$).
- ◇ A condição (ii) implica a limitação dos valores da matriz Hessiana (caso $f \in C^2$) e define M_2 (que corresponde no caso real a $\max |f''(x)|$).
- ◇ A condição (iii) permite garantir que as iteradas vão permanecer na bola $\bar{B}(x^{(0)}, \varepsilon_0)$ (e corresponde no caso real à condição $\left|\frac{f(a)}{f'(a)}\right| \leq b - a, \ \ \left|\frac{f(b)}{f'(b)}\right| \leq b - a$).
- ◇ A condição (iv) é óbvia e podemos mesmo considerar $\bar{D} = \bar{B}(x^{(0)}, \varepsilon_0)$.

Proposição (da convergência local). Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, zero de f, tal que $\det(J_f(z)) \neq 0$. Então o método de Newton converge para z desde que $x^{(0)}$ esteja suficientemente perto de z.

Proposição (da ordem de convergência). Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, zero de f, tal que $\det(J_f(z)) \neq 0$. Então o método de Newton, quando converge para z, tem convergência pelo menos quadrática, isto é, existe $K > 0$ tal que

$$\|z - x^{(m+1)}\| \leq K \|z - x^{(m)}|^2, \qquad m \in \mathbb{N},$$

ou

$$\|z - x^{(m)}\| \leq \frac{1}{K} \left(K \|z - x^{(0)}|\right)^{2^m}, \qquad m \in \mathbb{N}.$$

Exemplo. Considere o sistema de equações algébricas não-lineares

$$f(x) = 0, \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad f(x) = \begin{bmatrix} 3x_1 + x_2^2 \\ x_1^2 + 3x_2 - 1 \end{bmatrix}. \tag{S}$$

para o qual se verificou que tinha apenas duas raízes em $\mathbb{R}^2$, z e $\tilde{z}$.

(a) Determine o valor aproximado da raiz z usando três iteradas do método de Newton generalizado com aproximação inicial $x = [0 \ \ 0]^T$.

(b) Obtenha uma estimativa do erro da solução aproximada obtida na alínea (a) (usando a norma do máximo).

Resolução:

(a)

$$\begin{cases} x^{(m+1)} = x^{(m)} + \Delta x^{(m)} \\ J_f\left(x^{(m)}\right)\Delta x^{(m)} = -f\left(\tilde{x}^{(m)}\right) \end{cases}, \quad m \in \mathbb{N}, \qquad x^{(0)} = [0 \;\; 0]^T$$

$$J_f(x) = \begin{bmatrix} 3 & 2x_2 \\ 2x_1 & 3 \end{bmatrix}$$

m	$x_1^{(m)}$	$x_2^{(m)}$
0	0.000000000000000	0.000000000000000
1	0.000000000000000	0.333333333333333
2	−0.037037037037037	0.333333333333333
3	−0.036935981001465	0.332878581173261
4	−0.036936048808670	0.332878576099469
5	−0.036936048808670	0.332878576099468

(b)

$$J_f(x) = \begin{bmatrix} 3 & 2x_2 \\ 2x_1 & 3 \end{bmatrix}, \qquad [J_f(x)]^{-1} = \frac{1}{9 - 4x_1x_2}\begin{bmatrix} 3 & -2x_2 \\ -2x_1 & 3 \end{bmatrix}$$

$$H_{f_1}(x) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \qquad H_{f_2}(x) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\varepsilon_0 = 2\|\Delta x^{(0)}\|_\infty$$

$$\bar{D} = \bar{B}(x^{(0)}, \varepsilon_0) = [x_1^{(0)} - \varepsilon_0, x_1^{(0)} + \varepsilon_0] \times [x_2^{(0)} - \varepsilon_0, x_2^{(0)} + \varepsilon_0]$$

$$\frac{1}{M_1} = \max_{x \in D} \| \, [J_f(x)]^{-1} \, \|_\infty, \qquad \| \, [J_f(x)]^{-1} \, \|_\infty = \frac{\max\{3 + 2|x_2|, 3 + 2|x_1|\}}{|9 - 4x_1x_2|}$$

$$M_2 = \max_{x \in D}\{\|H_{f_1}(x)\|_\infty, \|H_{f_2}(x)\|_\infty\} = 2$$

$$K = \frac{M_2}{2M_1} = \frac{1}{M_1}, \qquad K\varepsilon_0 < \frac{1}{2}$$

$$\|z - x^{(m)}\|_\infty \leq \frac{1}{K}\left(K\varepsilon_0\right)^{2^m}$$

Estimativa de erro:

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \Delta x^{(0)} = \begin{bmatrix} 0 \\ \frac{1}{3} \end{bmatrix}$$

$$\varepsilon_0 = \frac{2}{3}, \qquad D = \left[-\frac{2}{3}, \frac{2}{3}\right] \times \left[-\frac{2}{3}, \frac{2}{3}\right]$$

$$K = \frac{1}{M_1} = \frac{3}{5} = 0.6, \qquad K\varepsilon_0 = \frac{2}{5} = 0.4 < \frac{1}{2}$$

$$\|z - x^{(3)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^8 = 0.109 \times 10^{-2}$$

$$\|z - x^{(4)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^{16} = 0.716 \times 10^{-6}$$

c.f. $\|z - x^{(3)}\|_\infty = 0.678 \times 10^{-7}, \qquad \|z - x^{(4)}\|_\infty = 0.1 \times 10^{-14}$

Melhoria da estimativa:

$$\tilde{x}^{(0)} = x^{(1)}) = \begin{bmatrix} 0 \\ \frac{1}{3} \end{bmatrix}, \qquad \Delta\tilde{x}^{(0)} = \begin{bmatrix} -\frac{1}{27} \\ 0 \end{bmatrix}$$

$$\varepsilon_0 = \frac{2}{27}, \qquad D = \left[-\frac{2}{27}, \frac{2}{27}\right] \times \left[\frac{7}{27}, \frac{11}{27}\right]$$

$$K = \frac{1}{M_1} = \frac{2781}{6473} = 0.429631, \qquad K\varepsilon_0 = \frac{206}{6473} = 0.0318245 < \frac{1}{2}$$

$$\|z - x^{(3)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^4 = 0.239 \times 10^{-5}$$

$$\|z - x^{(4)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^8 = 0.246 \times 10^{-11}$$

Melhoria da estimativa:

$$\tilde{x}^{(0)} = x^{(2)} = \begin{bmatrix} -\frac{1}{27} \\ \frac{1}{3} \end{bmatrix}, \qquad \Delta\tilde{x}^{(0)} = \begin{bmatrix} \frac{2}{19791} \\ -\frac{1}{2199} \end{bmatrix}$$

$$\varepsilon_0 = \frac{2}{2199}, \qquad D = \left[-\frac{751}{19791}, -\frac{715}{19791}\right] \times \left[\frac{731}{2199}, \frac{735}{2199}\right]$$

$$K = \frac{1}{M_1} = 0.405434, \qquad K\varepsilon_0 = 0.368744 \times 10^{-3} < \frac{1}{2}$$

$$\|z - x^{(3)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^2 = 0.336 \times 10^{-6}$$

$$\|z - x^{(4)}\|_\infty \le \frac{1}{K}\left(K\varepsilon_0\right)^4 = 0.457 \times 10^{-13}$$

6. INTERPOLAÇÃO POLINOMIAL

Introdução

Exemplo. Dada uma tabela de valores $\{(t_j, P_j), j = 0, 1, \ldots, n\}$, determinar um valor (aproximado) P_a correspondente a um valor $t_a \in]t_0, t_n[$.

Sendo k tal que $t_k < t_a < t_{k+1}$ façamos

$$x_0 = t_k, \qquad x_1 = t_{k+1}, \qquad x_2 = t_{k+2}$$
$$f_0 = P_k, \qquad f_1 = P_{k+1}, \qquad f_2 = P_{k+2}$$

Interpolação linear: $P_a \approx p_1(t_a)$, onde

$$p_1(x) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}\,(x - x_0)$$

Interpolação quadrática: $P_a \approx p_2(t_a)$, onde

$$p_2(x) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}\,(x - x_0) + \frac{\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}}{x_2 - x_0}\,(x - x_0)(x - x_1)$$

Problema. Sejam $x_0, x_1, \ldots, x_n$, $n+1$ pontos distintos do intervalo $[a, b]$ e sejam $f_0, f_1, \ldots, f_n$, os correspondentes valores de uma função $f : [a, b] \to \mathbb{R}$ nesses pontos,

$$f_j := f(x_j), \qquad j = 0, 1, \ldots, n.$$

Pretende-se determinar uma função $g : [a, b] \to \mathbb{R}$ pertencente a uma dada classe de funções tal que

$$g(x_j) = f_j, \qquad j = 0, 1, \ldots, n.$$

Os pontos x_j são chamados *pontos* ou *nós de interpolação* e os valores f_j os *valores interpolados*. A função g é chamada a *função interpoladora*. Vamos considerar funções interpoladoras polinomiais.

• Interessa saber se este problema tem solução, se a solução é única e como se calcula.

Proposição. Sejam $x_0, \ldots, x_n$, $n + 1$ pontos distintos do intervalo $[a, b]$ e sejam $f_0, \ldots, f_n$, os correspondentes valores de uma função $f : [a, b] \to \mathbb{R}$ nesses pontos. Existe um único polinómio p_n de grau menor ou igual a n que interpola f nos pontos $x_0, x_1, \ldots, x_n$, ou seja, tal que

$$p_n(x_j) = f_j, \qquad j = 0, 1, \ldots, n.$$

Dem.: $(\cdots)$

Notas.

(a) É importante distinguir entre a função p_n e as várias *representações* de p_n. Pelo teorema anterior sabemos que p_n é única para cada conjunto de $n+1$ pares ordenados $(x_j, f_j),\ j = 0, 1, \ldots, n$. Contudo poderá haver diversas maneiras de representar explicitamente p_n. A demonstração apresentada determina os coeficientes do polinómio por um sistema linear com matriz de Vandermonde. Já de seguida introduziremos a fórmula interpoladora de Lagrange. Mais adiante introduziremos a fórmula interpoladora de Newton. Cada uma dessas representações sugere um algoritmo para o cálculo de p_n.

(b) O facto do grau do polinómio interpolador poder ser menor ou igual a n tem a ver com a possibilidade dos valores $f_0, \ldots, f_n$, coincidirem com os valores de um polinómio q_m de grau $m \leq n$ nos nós de interpolação, isto é, $f_j = q_m(x_j),\ j = 0, 1, \ldots, n$.

(c) Há um número infinito de polinómios de grau $m > n$ que interpolam f nos pontos $x_0, \ldots, x_n$:

$$q_m(x) = p_n(x) + c \prod_{i=0}^{n} (x - x_i)^{\mu_i},$$

onde

$$\mu_i \in \mathbb{N}_1, \quad i = 0, 1, \ldots, n, \qquad m = \mu_0 + \mu_1 + \cdots + \mu_n.$$

Fórmula interpoladora de Lagrange

Proposição. Sejam $x_0, \ldots, x_n$, $n+1$ pontos distintos do intervalo $[a, b]$ e sejam $f_0, \ldots, f_n$, os correspondentes valores de uma função $f : [a, b] \to \mathbb{R}$ nesses pontos. O polinómio p_n de grau menor ou igual a n que interpola f nos pontos $x_0, x_1, \ldots, x_n$, é dado pela fórmula seguinte, conhecida por **fórmula interpoladora de Lagrange**:

$$p_n(x) = \sum_{j=0}^{n} f_j\, l_{j,n}(x), \qquad l_{j,n}(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}.$$

$l_{0,n}, l_{1,n}, \ldots, l_{n,n}$ são os *polinómios de Lagrange* de grau n associados aos nós $x_0, x_1, \ldots, x_n$.

Dem.: $(\cdots)$

Exemplo.

$$n = 1: \quad l_{0,1} = \frac{x - x_1}{x_0 - x_1}, \qquad l_{1,1} = \frac{x - x_0}{x_1 - x_0}\ ;$$

$$n = 2: \quad l_{0,2} = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \qquad l_{1,2} = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)},$$

$$l_{2,2} = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

Fórmula interpoladora de Newton

Proposição. Sejam $x_0, \ldots, x_n$, $n+1$ pontos distintos do intervalo $[a, b]$ e sejam $f_0, \ldots, f_n$, os correspondentes valores de uma função $f : [a, b] \to \mathbb{R}$ nesses pontos. O polinómio p_n de grau menor ou igual a n que interpola f nos pontos $x_0, x_1, \ldots, x_n$, é dado pela fórmula seguinte, conhecida por **fórmula interpoladora de Newton**:

$$p_n(x) = f[x_0] + \sum_{j=1}^{n} f[x_0, \ldots, x_j]\, W_j(x),$$

onde

$$W_j(x) = \prod_{i=0}^{j-1}(x - x_i), \qquad j = 1, 2, \ldots, n,$$

$$f[x_0] = f(x_0), \qquad f[x_0, \ldots, x_j] = \sum_{l=0}^{j} \frac{f(x_l)}{\prod\limits_{\substack{i=0 \\ i\neq l}}^{j}(x_l - x_i)}, \quad j = 1, 2, \ldots, n.$$

Os coeficientes $f[x_0, x_1, \ldots, x_j]$, $j = 0, 1, \ldots, n$, são conhecidos por **diferenças divididas de ordem j da função f associadas aos pontos** $x_0, x_1, \ldots, x_j$.

Dem.: $(\cdots)$

Exemplo.

$$f[x_0, x_1] = \frac{f[x_0]}{x_0 - x_1} + \frac{f[x_1]}{x_1 - x_0} = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_0]}{(x_0 - x_1)(x_0 - x_2)} + \frac{f[x_1]}{(x_1 - x_0)(x_1 - x_2)} + \frac{f[x_2]}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \end{aligned}$$

Proposição. Sendo $(i_0, i_1, \ldots, i_j)$ uma permutação de $(0, 1, \ldots, j)$ então

$$f[x_0, x_1, \ldots, x_j] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_j}].$$

Dem.: $(\cdots)$

Proposição.

$$f[x_0, x_1, \ldots, x_j] = \frac{f[x_1, x_2, \ldots, x_j] - f[x_0, x_1, \ldots, x_{j-1}]}{x_j - x_0}, \qquad j = 1, 2, 3, \ldots$$

Dem.: $(\cdots)$

• Estas fórmulas permitem calcular as diferenças divididas usando um esquema recursivo que é em geral apresentado sob a forma de uma tabela. Na $1^{\underline{a}}$ coluna escrevem-se os

pontos $x_0, x_1, \ldots$, e na $2^{\underline{a}}$ coluna os valores $f[x_0], f[x_1], \ldots$ Na $3^{\underline{a}}$ coluna escrevem-se as diferenças divididas de $1^{\underline{a}}$ ordem, obtidas por aplicação das fórmulas anteriores; e assim sucessivamente.

Tabela de diferenças divididas:

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$\cdots$
x_0	$f[x_0]$				
		$f[x_0, x_1]$			
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$		$\cdots$
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$		
		$f[x_3, x_4]$			
x_4	$f[x_4]$				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Nota.

$$p_2(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$
$$p_2(x) = f[x_1] + f[x_0, x_1](x - x_1) + f[x_0, x_1, x_2](x - x_1)(x - x_0)$$
$$p_2(x) = f[x_1] + f[x_1, x_2](x - x_1) + f[x_0, x_1, x_2](x - x_1)(x - x_2)$$
$$p_2(x) = f[x_2] + f[x_1, x_2](x - x_2) + f[x_0, x_1, x_2](x - x_2)(x - x_1)$$
$$p_2(x) = f[x_0] + f[x_0, x_2](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_2)$$
$$p_2(x) = f[x_2] + f[x_0, x_2](x - x_2) + f[x_0, x_1, x_2](x - x_2)(x - x_0)$$

Exemplo. Considere a seguinte tabela de valores de uma função f:

x_i	0	1	3	4
$f(x_i)$	0	2	8	9

Determine o polinómio interpolador de f nos pontos da tabela usando:

(a) O método da matriz de Vandermonde.

(b) A fórmula interpoladora de Lagrange.

(c) A fórmula interpoladora de Newton.

Resolução:

(a) Método da matriz de Vandermonde

$$p_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

$$p_3(x_j) = f_j, \quad j = 0, 1, 2, 3.$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 8 \\ 9 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 3 & 12 \\ 0 & 0 & 0 & 12 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \\ -3 \end{bmatrix}$$

$$a_3 = -\frac{1}{4}, \qquad a_2 = \frac{4}{3}, \qquad a_1 = \frac{11}{12}, \qquad a_0 = 0$$

$$p_3(x) = \frac{11}{12}\, x + \frac{4}{3}\, x^2 - \frac{1}{4} x^3 = \frac{x}{12}\left(-3x^2 + 16x + 11\right)$$

(b) Fórmula interpoladora de Lagrange:

$$p_3(x) = \sum_{j=0}^{3} f(x_j) l_j(x), \qquad l_j(x) = \prod_{i=0, i\neq j}^{3} \frac{x - x_i}{x_j - x_i}$$

$$p_3(x) = 2l_1(x) + 8l_2(x) + 9l_3(x)$$

$$l_1(x) = \frac{(x-0)(x-3)(x-4)}{(1-0)(1-3)(1-4)} = \frac{1}{6}\, x(x-3)(x-4)$$

$$l_2(x) = \frac{(x-0)(x-1)(x-4)}{(3-0)(3-1)(3-4)} = -\frac{1}{6}\, x(x-1)(x-4)$$

$$l_3(x) = \frac{(x-0)(x-1)(x-3)}{(4-0)(4-1)(4-3)} = \frac{1}{12}\, x(x-1)(x-3)$$

$$\begin{aligned} p_3(x) &= \frac{1}{3} x(x-3)(x-4) - \frac{4}{3} x(x-1)(x-4) + \frac{3}{4} x(x-1)(x-3)] \\ &= \frac{x}{12}\left(-3x^2 + 16x + 11\right) \end{aligned}$$

(c) Fórmula interpoladora de Newton às diferenças divididas:

$$\begin{aligned} p_3(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\quad + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \end{aligned}$$

i	x_i	$f[x_i]$	$f[\cdot,\cdot]$	$f[\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\cdot,\cdot]$
0	0	0			
			2		
1	1	2		$\frac{1}{3}$	
			3		$-\frac{1}{4}$
2	3	8		$-\frac{2}{3}$	
			1		
3	4	9		1	

$$p_3(x) = 2x + \frac{1}{3}x(x-1) - \frac{1}{4}x(x-1)(x-3) = \frac{x}{12}\left(-3x^2 + 16x + 11\right)$$

Erro de interpolação polinomial

Proposição. Seja p_n o polinómio interpolador de f em $n+1$ pontos distintos $x_0, x_1, \ldots, x_n$ de $[a,b]$. Se $f \in C^{n+1}([a,b])$ então para cada $x \in [a,b]$ existe um ponto $\xi(x) \in]x_0; x_1; \ldots; x_n; x[$ tal que

$$e_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\, W_{n+1}(x),$$

onde

$$W_{n+1}(x) = \prod_{i=0}^{n}(x - x_i).$$

Dem.: $(\cdots)$

Proposição. Seja p_n o polinómio interpolador de f em $n+1$ pontos distintos $x_0, x_1, \ldots, x_n$ de $[a,b]$. Se $f \in C^{n+1}([a,b])$ então

$$e_n(x) = f[x_0, x_1, \ldots, x_n, x] W_{n+1}(x).$$

Dem.: $(\cdots)$

Proposição. Seja $f \in C^{n+1}([a,b])$ e sejam $x_0, x_1, \ldots, x_n$, $n+1$ pontos distintos de $[a,b]$. Então:

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}, \qquad \xi \in]x_0; x_1; \ldots; x_n[.$$

Dem.: $(\cdots)$

• A equação que obtivemos para o erro de interpolação de Lagrange não pode ser usada para calcular o valor exacto do erro $e_n(x)$ visto que $\xi(x)$ é em geral uma função desconhecida. Uma excepção é o caso em que $f^{(n+1)}$ é uma constante. Pode no entanto ser usada para obter um majorante do erro.

Proposição. Sendo $f \in C^{n+1}([a,b])$ então:

(1)
$$|e_n(x)| \leq \frac{M_{n+1}}{(n+1)!}\,|W_{n+1}(x)|, \qquad x \in [a,b],$$

onde
$$M_{n+1} = \max_{x\in[a,b]} |f^{(n+1)}(x)|.$$

(2)
$$|e_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \max_{\tilde{x}\in[a,b]} |W_{n+1}(\tilde{x})|, \qquad \forall x \in [a,b].$$

Dem.: $(\cdots)$

• Estas fórmulas pôem em evidência a importância da contribuição da função $|W_{n+1}|$ para o erro de interpolação. Vamos estudar em mais algum pormenor esta função distinguindo dois casos: nós igualmente espaçados e nós de Chebyshev.

Proposição. Suponhamos que os nós de interpolação estão igualmente espaçados entre si:

$$x_i = x_0 + ih, \quad i = 0, 1, \ldots, n,$$

onde $h > 0$ é tal que $x_n \leq b$. Sendo $f \in C^{n+1}([a,b])$ a fórmula de majoração do erro de interpolação escreve-se

$$|e_n(x)| \leq \frac{h^{n+1} M_{n+1}}{(n+1)!}\,|\Psi_{n+1}(t)|, \qquad \forall x \in [a,b],$$

onde $x = x_0 + th$, e
$$\Psi_{n+1}(t) = \prod_{i=0}^{n} (t-i).$$

Dem.: $(\cdots)$

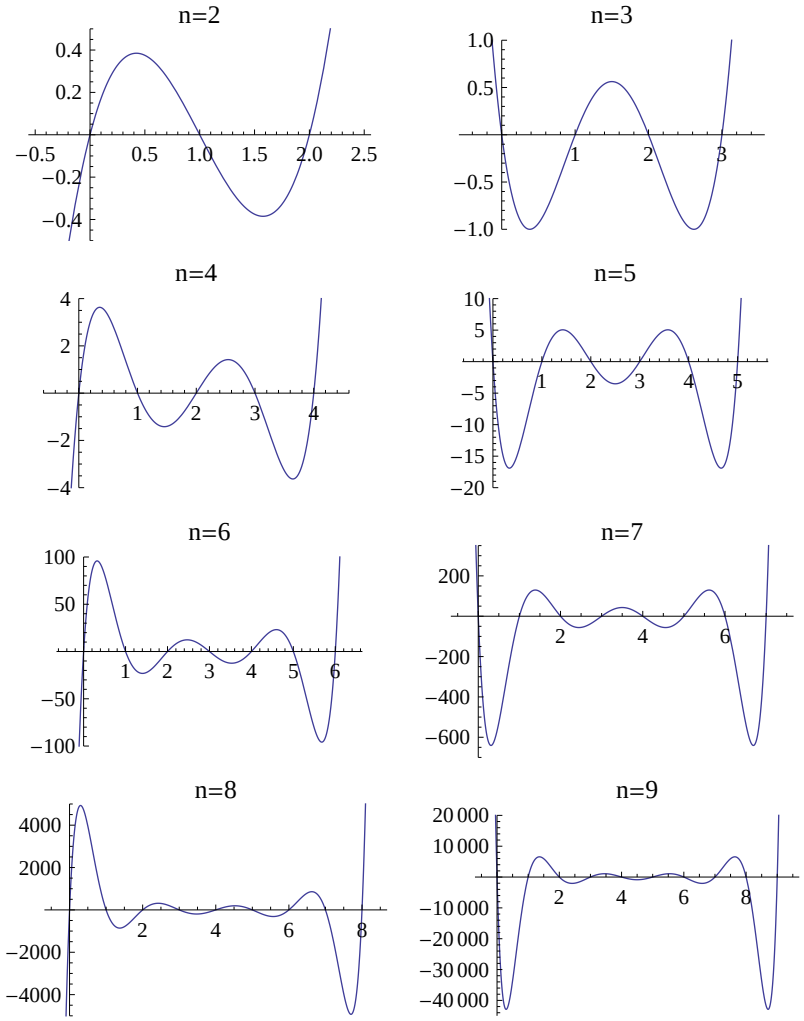
Proposição. A função Ψ_{n+1} tem as seguintes propriedades (ilustradas na figura seguinte):

(1) Ψ_{n+1} é um polinómio mónico de grau $n+1$ com zeros nos pontos $0, 1, \ldots, n$.

(2) Ψ_{n+1} é uma função simétrica ou antisimétrica em relação ao ponto médio dos zeros, $t = \frac{n}{2}$, consoante $n+1$ é par ou ímpar, respectivamente, isto é,
$$\Psi_{n+1}\left(\frac{n}{2} - t\right) = (-1)^{n+1} \Psi_{n+1}\left(\frac{n}{2} + t\right).$$

(3) Em cada intervalo $]i, i+1[$, $i = 0, 1, \ldots, n-1$, $|\Psi_{n+1}|$ possui um único máximo.

(4) Os valores dos máximos relativos de $|\Psi_{n+1}|$ crescem à medida que t se afasta de $\frac{n}{2}$.



(5) $|\Psi_{n+1}|$ tem um máximo absoluto no intervalo $[0, n]$, o qual ocorre em dois pontos, um no subintervalo $]0, 1[$ e outro no subintervalo $]n-1, n[$, verificando-se a desigualdade

$$\max_{t\in[0,n]} |\Psi_{n+1}(t)| < n!.$$

(6) Este valor máximo absoluto e o seu quociente em relação aos valores dos restantes máximos relativos crescem com n.

(7) $|\Psi_{n+1}(t)|$ cresce rapidamente para $t < 0$ e $t > n$.

Notas. Atendendo à fórmula de majoração do erro e ao que se disse sobre a função Ψ_{n+1} conclui-se que:

(a) na interpolação de f por p_n os nós de interpolação igualmente espaçados $x_0, x_0 + h, \ldots, x_0 + nh$, devem ser escolhidos por forma a que x esteja perto do centro do intervalo $[x_0, x_0 + nh]$, isto é, $x \approx x_0 + \frac{nh}{2}$.

(b) o erro de interpolação cresce muito rapidamente quando x se afasta de x_0, para a esquerda, e de $x_0 + nh$, para a direita, o que desencoraja a *extrapolação*, isto é, a aproximação de $f(x)$ por $p_n(x)$ para valores de x fora do intervalo $[x_0, x_0 + nh]$.

Proposição. De entre todas as escolhas de nós de interpolação $x_0, x_1, \ldots, x_n$ distintos no intervalo $[-1, 1]$, a quantidade

$$\max_{x\in[-1,1]} \left| \prod_{i=0}^{n} (x - x_i) \right|,$$

toma o menor valor possível para

$$x_k = -\cos\left(\frac{2k+1}{2n+2}\,\pi\right), \qquad k = 0, 1, \ldots, n,$$

e esse valor é 2^{-n}. Se $f \in C^{n+1}([-1, 1])$ a fórmula de majoração do erro de interpolação escreve-se

$$|e_n(x)| \leq \frac{M_{n+1}}{(n+1)!2^n}, \qquad \forall x \in [-1, 1].$$

Nota. Os nós que acabámos de introduzir são chamados **nós de Chebyshev**. Eles são os $n + 1$ zeros do chamado **polinómio de Chebyshev** de grau $n + 1$, habitualmente designado por T_{n+1}, que introduziremos no capítulo seguinte.

Exemplo. Considere a função $f : \mathbb{R} \to \mathbb{R}$ definida por $f(x) = \cos x$.

(a) Determine o polinómio interpolador de f nos pontos $x_0 = -a,\ x_1 = 0, x_2 = a$, onde $a \in]0, 1]$.

(b) Determine um majorante do erro de interpolação válido para todos os pontos do intervalo $[-1, 1]$.

(c) Determine o valor de a para o qual o majorante do erro de interpolação tem o menor valor possível.

Resolução:

(a) $p_2(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$

x_i	$f[x_i]$	$f[\cdot,\cdot]$	$f[\cdot,\cdot,\cdot]$
$-a$	$\cos a$		
		$\dfrac{1-\cos a}{a}$	
0	1		$\dfrac{\cos a - 1}{a^2}$
		$\dfrac{\cos a - 1}{a}$	
a	$\cos a$		

$$\begin{aligned} p_2(x) &= \cos a + \frac{1-\cos a}{a}\,(x + a) + \frac{\cos a - 1}{a^2}\,(x + a)x \\ &= 1 - \frac{1-\cos a}{a^2}\,x^2 \end{aligned}$$

(b) $e_2(x) = \dfrac{f'''(\xi)}{6}\,W_3(x)$

$$|e_2(x)| \leq \frac{1}{6}\,\|f'''\|_\infty \|W_3\|_\infty$$

$$\|f'''\|_\infty = \max_{x\in[-1,1]} |f'''(x)| = \max_{x\in[-1,1]} |\sin x| = \sin(1)$$

$$\|W_3\|_\infty = \max_{x\in[-1,1]} |W_3(x)|$$

$$W_3(x) = (x+a)x(x-a) = x(x^2-a^2)$$

$$W_3'(x) = 3x^2 - a^2$$

$$\|W_3\|_\infty = \max\left\{W_3\left(-\frac{a}{\sqrt{3}}\right), W_3(1)\right\} = \max\left\{\frac{2a^3}{3\sqrt{3}}, 1-a^2\right\}$$

$$= \begin{cases} 1-a^2, & a \leq \dfrac{\sqrt{3}}{2} \\ \dfrac{2a^3}{3\sqrt{3}}, & a \geq \dfrac{\sqrt{3}}{2} \end{cases}$$

(c) $\min\limits_{a\in]0,1]} \|W_3\|_\infty = \dfrac{1}{4} \qquad \left(\text{para } a = \dfrac{\sqrt{3}}{2}\right)$

Note-se que $-\dfrac{\sqrt{3}}{2},\ 0,\ \dfrac{\sqrt{3}}{2}$ são os nós de Chebyshev para $n = 2$.

7. APROXIMAÇÃO MÍNIMOS QUADRADOS

Introdução

• É muitas vezes de interesse o problema de aproximar uma função "complicada" por uma função "simples". No capítulo anterior considerámos o caso em que as funções aproximadoras são os polinómios interpoladores.

Neste capítulo consideraremos outros tipos de funções aproximadoras para f, sem a exigência que elas interpolem f. Com efeito, em muitas situações não é conveniente que a função aproximadora seja uma função interpoladora. Por exemplo, suponhamos que a função f traduz uma relação entre duas grandezas físicas, x e $f(x)$. Suponhamos ainda que os valores $f(x_i),\ i = 0, 1, \ldots, n$, foram obtidos experimentalmente; ou seja, apenas dispomos de valores $f(x_i) + \varepsilon_i,\ i = 0, 1, \ldots, n$, onde os erros ε_i são desconhecidos. Neste caso, não seria razoável aproximar f por um polinómio passando por $(x_i, f(x_i) + \varepsilon_i),\ i = 0, 1, \ldots, n$, a não ser que os erros ε_i fossem pequenos. Uma solução mais adequada seria utilizar um polinómio obtido pelo chamado **método dos mínimos quadrados**, onde se reduz o efeito dos erros dos dados.

• Para se avaliar a qualidade de uma aproximação precisamos de introduzir uma noção de distância entre a função e a função aproximadora.

Definição. Seja E um conjunto não vazio. Uma função $d : E \times E \to \mathbb{R}$ que verifica as condições

(M1) $d(f, g) \geq 0$ e $d(f, g) = 0 \ \Leftrightarrow \ f = g, \ \forall f, g \in E$;
(M2) $d(f, g) = d(g, f), \ \forall f, g \in E$;
(M3) $d(f, g) \leq d(f, h) + d(h, g), \ \forall f, g, h \in E$;

diz-se uma **métrica** ou **distância**. O conjunto E onde está definida a métrica diz-se um **espaço métrico**.

Problema geral da aproximação. Seja E um espaço métrico e $F \subset E$ um subespaço de E. Dado um elemento $f \in E$ pretende saber-se se existe um elemento $\phi^* \in F$ tal que

$$d(f, \phi^*) = \inf_{\phi \in F} d(f, \phi).$$

Se existir então ϕ^* é chamada uma **melhor aproximação (m.a.)** de f em F relativamente à métrica ou distância d. Se existir interessa saber se é único, quais as suas propriedades e como pode ser construído.

• Dado $f \in E$, o problema de determinar a m.a. ϕ^* de f em $F \subset E$ é em geral difícil. Vamos apenas considerar o importante caso em que E é um espaço pré-Hilbertiano, isto é, um espaço vectorial munido de um produto interno. Veremos que a determinação da m.a. relativamente à distância induzida pela norma induzida pelo produto interno pode ser reduzida à resolução de um sistema de equações lineares.

Melhor aproximação em espaços pré-Hilbertianos

Definição. Seja E um espaço vectorial. Uma função $\langle\ ,\ \rangle : E \times E \to \mathbb{R}$, que verifica as condições

(P1) $\langle f, f\rangle \geq 0$ e $\langle f, f\rangle = 0 \ \Leftrightarrow \ f = 0, \ \forall f \in E$;
(P2) $\langle f, g\rangle = \langle g, f\rangle, \ \forall f, g \in E$;
(P3) $\langle f, \alpha g + \beta h\rangle = \alpha\langle f, g\rangle + \beta\langle f, h\rangle, \ \forall f, g, h \in E, \ \forall \alpha, \beta \in \mathbb{R}$;

diz-se um **produto interno**. O espaço E onde está definido um produto interno diz-se um **espaço pré-Hilbertiano (p-H)**.

• Um produto interno induz a norma

$$\|f\| = \sqrt{\langle f, f\rangle},$$

passando E a ser um espaço normado. Uma norma induz uma distância

$$d(f, g) = \|f - g\|,$$

passando E a ser um espaço métrico.

Exemplo. $E = \mathbb{R}^d$ dotado do produto interno

$$\langle f, g\rangle = \sum_{i=1}^{d} w_i f_i g_i, \qquad \forall f, g \in E,$$

onde $f = (f_1, f_2, \ldots f_d), g = (g_1, g_2, \ldots g_d)$ e $w_1, w_2, \ldots, w_d$ são constantes positivas, é um espaço p-H.

Exemplo. $E = C([a, b])$ dotado do produto interno

$$\langle f, g\rangle = \int_a^b w(x) f(x) g(x)\, dx, \qquad \forall f, g \in C([a, b]), \tag{\#}$$

onde w é uma função de peso, é um espaço p-H.

Definição. Uma função $w : [a, b] \to \mathbb{R}$ é uma **função de peso** se verificar as seguintes condições:

(i) $w(x) \geq 0, \quad \forall x \in [a, b]$;
(ii) $\{x \in [a, b] : w(x) = 0\}$ é finito;
(iii) $\displaystyle\int_a^b x^n w(x) dx$ existe e é finito para qualquer $n \in \mathbb{N}_0$.

Exemplos.

(i) $w(x) = 1, \quad x \in [a, b]$;

(ii) $w(x) = \dfrac{1}{\sqrt{1-x^2}}, \quad x \in]-1,1[;$

(iii) $w(x) = e^{-x}, \quad x \in [0,\infty[\ ;$

(iv) $w(x) = e^{-x^2}, \quad x \in]-\infty,+\infty[.$

Problema I. Sejam $E = C([a,b])$ e F um subespaço de E de dimensão finita. Dada $f \in E$ determinar $\phi^* \in F$ que minimiza o integral

$$\int_a^b w(x)[f(x)-\phi(x)]^2\,dx.$$

Chama-se a ϕ^* a **melhor aproximação mínimos quadrados (m.a.m.q.)** de f em F.

Nota. Em termos do produto interno $(\#)$ em $C([a,b])$ o problema toma a forma: dada $f \in E$ determinar $\phi^* \in F$ tal que

$$\|f-\phi^*\| = \min_{\phi\in F} \|f-\phi\|.$$

Problema II. Sejam $E = C([a,b])$, F um subespaço de E de dimensão finita $n+1$ e $\Pi = \{x_0, x_1, \ldots, x_N\}$ um conjunto de $N+1$ pontos distintos de $[a,b]$, verificando-se que $N \geq n$. Dada $f \in E$ determinar $\phi^* \in F$ que minimiza a soma

$$\sum_{i=0}^{N} w_i[f(x_i)-\phi(x_i)]^2.$$

Chama-se a ϕ^* a **melhor aproximação mínimos quadrados discreta (m.a.m.q.d.)** de f em F.

Nota. Em termos do produto interno em $\mathbb{R}^d$ acima definido o problema toma a seguinte forma. Começa-se por associar a cada $f \in E$ um vector $\bar{f} \in \mathbb{R}^{N+1}$ tal que $\bar{f}_i = f(x_i), i \in \{0,1,\ldots,N\}$. Seja ainda $\bar{F} = \{\bar{\phi} \in \mathbb{R}^{N+1} : \phi \in F\} \subset \mathbb{R}^{N+1}$. Dada $f \in E$ determinar $\bar{\phi}^* \in \bar{F}$ tal que

$$\|\bar{f}-\bar{\phi}^*\| = \min_{\bar{\phi}\in\bar{F}} \|\bar{f}-\bar{\phi}\|.$$

Proposição (Caracterização da m.a. num espaço p-H). Sejam E um espaço p-H e F um subespaço de E. Dado $f \in E$, o elemento $\phi^* \in F$ é uma m.a. de f em F se e só se

$$\langle f-\phi^*, \phi\rangle = 0, \qquad \forall \phi \in F.$$

A m.a. de f em F é única.

Dem.: $(\cdots)$

Proposição (Sistema normal). Sejam E um espaço p-H, F um subespaço de E de dimensão finita $n+1$ e $\{\varphi_0,\ldots,\varphi_n\}$ uma base de F. Então a m.a. ϕ^* de $f \in E$ em F é dada por

$$\phi^* = \sum_{k=0}^{n} a_k^* \varphi_k,$$

onde $a_0^*, \ldots, a_n^*$ é a solução única do **sistema normal**

$$\sum_{k=0}^{n} \langle \varphi_j, \varphi_k \rangle a_k^* = \langle f, \varphi_j \rangle, \qquad j = 0, 1, \ldots, n,$$

ou

$$\begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \cdots & \langle \varphi_0, \varphi_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \varphi_n, \varphi_0 \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \begin{bmatrix} a_0^* \\ \vdots \\ a_n^* \end{bmatrix} = \begin{bmatrix} \langle f, \varphi_0 \rangle \\ \vdots \\ \langle f, \varphi_n \rangle \end{bmatrix}.$$

Exemplo. Determinar a m.a.m.q. de $f \in E = C([0,1])$ no subespaço F gerado pela base $\{\varphi_0, \ldots \varphi_n\}$, $\varphi_j(x) = x^j$, no caso em que $w(x) = 1, \forall x \in [0,1]$. Note-se que a matriz do sistema normal é a matriz de Hilbert de ordem $n+1$.

Resolução:

$$\phi^* = \sum_{k=0}^{n} a_k^* \varphi_k, \qquad \varphi_k(x) = x^k$$

$$\begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \cdots & \langle \varphi_0, \varphi_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \varphi_n, \varphi_0 \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \begin{bmatrix} a_0^* \\ \vdots \\ a_n^* \end{bmatrix} = \begin{bmatrix} \langle f, \varphi_0 \rangle \\ \vdots \\ \langle f, \varphi_n \rangle \end{bmatrix}.$$

$$\langle \varphi_j, \varphi_k \rangle = \int_0^1 x^{j+k} dx = \frac{1}{j+k+1} = (H_{n+1})_{jk}$$

$$\langle f, \varphi_j \rangle = \int_0^1 x^j f(x) dx$$

Exemplo. Determinar a m.a.m.q.d. de $f \in E = C([a,b])$ no subespaço F gerado pela base $\{\varphi_0, \ldots \varphi_n\}$, no caso em que $w_i = 1$, $\forall i \in \{0, 1, \ldots, n\}$. Particularizar para o caso em que $\varphi_j(x) = x^j$.

Resolução:

$$\phi^* = \sum_{k=0}^{n} a_k^* \varphi_k, \qquad \bar{\phi}^* = \sum_{k=0}^{n} a_k^* \bar{\varphi}_k$$

$$\begin{bmatrix} \langle \bar{\varphi}_0, \bar{\varphi}_0 \rangle & \cdots & \langle \bar{\varphi}_0, \bar{\varphi}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \bar{\varphi}_n, \bar{\varphi}_0 \rangle & \cdots & \langle \bar{\varphi}_n, \bar{\varphi}_n \rangle \end{bmatrix} \begin{bmatrix} a_0^* \\ \vdots \\ a_n^* \end{bmatrix} = \begin{bmatrix} \langle \bar{f}, \bar{\varphi}_0 \rangle \\ \vdots \\ \langle \bar{f}, \bar{\varphi}_n \rangle \end{bmatrix}.$$

$$\bar{\varphi}_j = [\varphi_j(x_0) \;\; \varphi_j(x_1) \ldots \varphi_j(x_N)]^T, \qquad \bar{f} = [f(x_0) \;\; f(x_1) \ldots f(x_N)]^T$$

$$\langle \bar{\varphi}_j, \bar{\varphi}_k \rangle = \sum_{i=0}^{N} \varphi_j(x_i) \varphi_k(x_i), \qquad \langle \bar{f}, \bar{\varphi}_j \rangle = \sum_{i=0}^{N} f(x_i) \varphi_j(x_i)$$

Caso particular: $\varphi_j(x) = x^j$

$$\langle \bar{\varphi}_j, \bar{\varphi}_k \rangle = \sum_{i=0}^{N} x_i^{j+k}, \qquad \langle \bar{f}, \bar{\varphi}_j \rangle = \sum_{i=0}^{N} f(x_i) x_i^j$$

Note-se que $\langle \bar{\varphi}_0, \bar{\varphi}_0 \rangle = N + 1$

Sistemas ortogonais.

Definição. Um conjunto S de elementos de um espaço p-H E diz-se **ortogonal** se

$$\langle u, v \rangle = 0, \quad \forall u, v \in S, \quad u \neq v,$$

(e usa-se a notação $u \perp v$). Se, além disso,

$$\langle u, u \rangle = 1, \quad \forall u \in S,$$

o sistema diz-se **ortonormado**.

Proposição (Teorema de Pitágoras). Sendo $\{u_0, u_1, \ldots, u_n\}$ um conjunto ortogonal então

$$\left\| \sum_{i=0}^{n} u_i \right\|^2 = \sum_{i=0}^{n} \|u_i\|^2.$$

Dem.: $(\cdots)$

Proposição. Um conjunto finito ortogonal de elementos não nulos é um conjunto de elementos linearmente independentes.

Dem.: $(\cdots)$

Proposição (Ortogonalização de Gram-Schmidt). Seja $\{e_0, \ldots, e_n\}$ uma base de um subespaço F de um espaço p-H E. Seja $\{\varphi_0, \ldots, \varphi_n\}$ o conjunto de elementos definidos por

$$\varphi_0 = e_0, \qquad \varphi_j = e_j - \sum_{k=0}^{j-1} \frac{\langle e_j, \varphi_k \rangle}{\langle \varphi_k, \varphi_k \rangle} \varphi_k, \quad j \geq 1,$$

e seja $\{\hat{\varphi}_0, \ldots, \hat{\varphi}_n\}$ o conjunto de elementos definidos por

$$\hat{\varphi}_j = \frac{\varphi_j}{\|\varphi_j\|}.$$

Então o conjunto $\{\varphi_0, \ldots, \varphi_n\}$ constitui uma base ortogonal para F enquanto que o conjunto $\{\hat{\varphi}_0, \ldots, \hat{\varphi}_n\}$ constitui uma base ortonormada para F.

Exemplo. Determinar o sistema de polinómios ortogonais no intervalo $[-1, 1]$ em relação ao produto interno definido por $(\#)$ com $w(x) = 1$, $\forall x \in [-1, 1]$.

Resolução:

$$e_j(x) = x^j, \quad j \geq 0$$

$$\varphi_0 = e_0, \qquad \varphi_0(x) = 1$$

$$\varphi_1 = e_1 - \frac{\langle e_1, \varphi_0 \rangle}{\langle \varphi_0, \varphi_0 \rangle} \varphi_0$$

$$\langle e_1, \varphi_0 \rangle = \int_{-1}^{1} x \, dx = 0$$

$$\varphi_1 = e_1, \qquad \varphi_1(x) = x$$

$$\varphi_2 = e_2 - \frac{\langle e_2, \varphi_0 \rangle}{\langle \varphi_0, \varphi_0 \rangle} \varphi_0 - \frac{\langle e_2, \varphi_1 \rangle}{\langle \varphi_1, \varphi_1 \rangle} \varphi_1$$

$$\langle e_2, \varphi_0 \rangle = \int_{-1}^{1} x^2 \, dx = \frac{2}{3}, \qquad \langle \varphi_0, \varphi_0 \rangle = \int_{-1}^{1} dx = 2$$

$$\langle e_2, \varphi_1 \rangle = \int_{-1}^{1} x^3 \, dx = 0$$

$$\varphi_2 = e_2 - \frac{1}{3} \varphi_0, \qquad \varphi_2(x) = x^2 - \frac{1}{3}$$

$$\varphi_3 = e_3 - \frac{\langle e_3, \varphi_0 \rangle}{\langle \varphi_0, \varphi_0 \rangle} \varphi_0 - \frac{\langle e_3, \varphi_1 \rangle}{\langle \varphi_1, \varphi_1 \rangle} \varphi_1 - \frac{\langle e_3, \varphi_2 \rangle}{\langle \varphi_2, \varphi_2 \rangle} \varphi_2$$

$$\langle e_3, \varphi_0 \rangle = \int_{-1}^{1} x^3 \, dx = 0$$

$$\langle e_3, \varphi_1 \rangle = \int_{-1}^{1} x^4 \, dx = \frac{2}{5}, \qquad \langle \varphi_1, \varphi_1 \rangle = \int_{-1}^{1} x^2 \, dx = \frac{2}{3}$$

$$\langle e_3, \varphi_2 \rangle = \int_{-1}^{1} x^3 \left(x^2 - \frac{1}{3} \right) dx = 0$$

$$\varphi_3 = e_3 - \frac{3}{5} \varphi_1, \qquad \varphi_3(x) = x^3 - \frac{3}{5} x$$

Proposição (M.a. num espaço p-H usando uma base ortogonal (ortonormada)). Sendo E um espaço p-H, F um subespaço de E de dimensão finita e $\{\varphi_O, \ldots, \varphi_n\}$ uma base ortogonal de F, então a m.a. ϕ^* de $f \in E$ em F é dada por

$$\phi^* = \sum_{k=0}^{n} a_k^* \varphi_k, \qquad a_k^* = \frac{\langle f, \varphi_k \rangle}{\langle \varphi_k, \varphi_k \rangle}.$$

Se a base for ortonormada ter-se-á simplesmente

$$\phi^* = \sum_{k=0}^{n} a_k^* \varphi_k, \qquad a_k^* = \langle f, \varphi_k \rangle.$$

Dem.: $(\cdots)$

Polinómios ortogonais.

• Consideremos o espaço $C([a,b])$ com o produto interno definido por (#). O processo de ortogonalização de Gram-Schmidt aplicado ao conjunto $\{1, x, x^2, \ldots, x^n\}$ permite obter diferentes bases ortogonais para $\mathcal{P}_n$, o conjunto de polinómios de grau menor ou igual a n, correspondentes a diferentes funções de peso w. Uma maneira alternativa de construir polinómios ortogonais é usar o teorema seguinte.

Proposição (Relação de recorrência tripla). O conjunto de polinómios $\{\varphi_0, \ldots, \varphi_n\}$, definidos por

$$\varphi_0(x) = 1, \qquad \varphi_1(x) = x - B_1,$$

$$\varphi_k(x) = (x - B_k)\varphi_{k-1}(x) - C_k\varphi_{k-2}(x), \qquad k \geq 2,$$

com

$$B_k = \frac{\langle x\varphi_{k-1}, \varphi_{k-1}\rangle}{\langle \varphi_{k-1}, \varphi_{k-1}\rangle}, \qquad k \geq 1, \qquad\qquad C_k = \frac{\langle x\varphi_{k-1}, \varphi_{k-2}\rangle}{\langle \varphi_{k-2}, \varphi_{k-2}\rangle}, \qquad k \geq 2,$$

é ortogonal em relação ao produto interno definido por (#), isto é,

$$\langle \varphi_j, \varphi_k\rangle = \int_a^b w(x)\varphi_j(x)\varphi_k(x)\,dx = 0, \qquad j \neq k.$$

Nota. Esta via produz polinómios ortogonais **mónicos**, isto é, com coeficente unitário da potência mais elevada.

Exemplo. Determinar o sistema de polinómios ortogonais no intervalo $[-1, 1]$ em relação ao produto interno definido por (#) com $w(x) = 1,\ \forall x \in [-1, 1]$.

Resolução:

$\varphi_0(x) = 1$

$$B_1 = \frac{\langle x\varphi_0, \varphi_0\rangle}{\langle \varphi_0, \varphi_0\rangle} = \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} dx} = 0$$

$\varphi_1(x) = x$

$$B_2 = \frac{\langle x\varphi_1, \varphi_1\rangle}{\langle \varphi_1, \varphi_1\rangle} = \frac{\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} x^2\,dx} = 0$$

$$C_2 = \frac{\langle x\varphi_1, \varphi_0\rangle}{\langle \varphi_0, \varphi_0\rangle} = \frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} dx} = \frac{\frac{2}{3}}{2} = \frac{1}{3}$$

$\varphi_2(x) = x^2 - \dfrac{1}{3}$

$$B_3 = \frac{\langle x\varphi_2, \varphi_2\rangle}{\langle \varphi_2, \varphi_2\rangle} = \frac{\int_{-1}^{1} x^5\,dx}{\int_{-1}^{1} x^4\,dx} = 0$$

$$C_3 = \frac{\langle x\varphi_2, \varphi_1\rangle}{\langle \varphi_1, \varphi_1\rangle} = \frac{\int_{-1}^{1} x^2\left(x^2 - \frac{1}{3}\right)\,dx}{\int_{-1}^{1} x^2\,dx} = \frac{\frac{8}{45}}{\frac{2}{3}} = \frac{4}{15}$$

$$\varphi_3(x) = x^3 - \frac{3}{5}\,x$$

Proposição. Seja $\{\varphi_0, \ldots, \varphi_n\}$ um sistema de polinómios ortogonais em relação ao produto n, interno definido por (#). Admite-se implicitamente que grau$(\varphi_j) = j$. Se p é um polinómio de grau $m < n$ então:

$$(1)\ \ \langle p, \varphi_j\rangle = 0, \quad j = m+1, \ldots, n; \qquad (2)\ \ p = \sum_{j=0}^{m} \frac{\langle p, \varphi_j\rangle}{\langle \varphi_j, \varphi_j\rangle}\,\varphi_j.$$

Proposição. Seja $\{\varphi_0, \ldots, \varphi_n\}$ um sistema de polinómios ortogonais em relação ao produto interno definido por (#). Admite-se implicitamente que grau$(\varphi_j) = j$. Então o polinómio φ_j tem exactamente j raízes reais distintas no intervalo aberto $]a, b[$.

Exemplo. Polinómios de Legendre, P_n $(x \in [a,b] = [-1,1],\ \ w(x) = 1)$:

$$\begin{cases} P_n(x) = \dfrac{2n-1}{n}\,xP_{n-1}(x) - \dfrac{n-1}{n}\,P_{n-2}(x), & n \geq 2 \\ P_0(x) = 1, \qquad P_1(x) = x \end{cases}$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1), \qquad P_3(x) = \frac{1}{2}(5x^3 - 3x), \qquad P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

$$\text{grau}(P_n) = n, \qquad P_n(1) = 1, \qquad n \geq 0$$

$$A_n = \lim_{x\to\infty} x^{-n}P_n(x) = \frac{(2n)!}{2^n(n!)^2}, \qquad n \geq 1$$

$$\langle P_m, P_n\rangle = \int_{-1}^{+1} P_m(x)P_n(x)\,dx = \begin{cases} 0, & m \neq n, \\ \dfrac{2}{2n+1}, & m = n. \end{cases}$$

Exemplo. Polinómios de Chebyshev, T_n $\left(x \in [a,b] = [-1,1],\ \ w(x) = \dfrac{1}{\sqrt{1-x^2}}\right)$:

$$\begin{cases} T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), & n \geq 2 \\ T_0(x) = 1, \qquad T_1(x) = x \end{cases}$$

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x, \qquad T_4(x) = 8x^4 - 8x^2 + 1$$

$$\text{grau}(T_n) = n, \qquad T_n(1) = 1, \qquad n \geq 0$$

$$A_n = \lim_{x\to\infty} x^{-n} T_n(x) = 2^{n-1}, \qquad n \geq 1$$

$$\langle T_m, T_n \rangle = \int_{-1}^{+1} \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}}\, dx = \begin{cases} 0, & m \neq n, \\ \pi, & m = n = 0, \\ \dfrac{\pi}{2}, & m = n > 0 \end{cases}$$

$$T_n(x) = \cos(n \arccos x), \qquad n \geq 0$$

$$T_n(x_i) = 0, \qquad x_i = -\cos\frac{(2i+1)\pi}{2n}, \qquad i = 0, \ldots, n-1, \qquad n \geq 1$$

Exemplo. Determine de entre os polinómios de grau menor ou igual a 2 a m.a.m.q. da função $f : \mathbb{R} \to \mathbb{R}$, definida por $f(x) = x^4 + x^3$, relativamente aos seguintes produtos internos:

(a) $\langle g, h \rangle = \displaystyle\int_{-1}^{1} g(x) h(x)\, dx, \qquad \forall g, h \in C([a,b]).$

Sugestão: utilize os polinómios de Legendre.

(b) $\langle g, h \rangle = \displaystyle\int_{-1}^{1} \frac{g(x)h(x)}{\sqrt{1-x^2}}\, dx, \qquad \forall g, h \in C([a,b]).$

Sugestão: utilize os polinómios de Chebyshev.

Resolução:

(a)

$$\begin{aligned} P_0(x) &= 1 & 1 &= P_0(x) \\ P_1(x) &= x & x &= P_1(x) \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) & x^2 &= \frac{1}{3}[P_0(x) + 2P_2(x)] \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) & x^3 &= \frac{1}{5}[3P_1(x) + 2P_3(x)] \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) & x^4 &= \frac{1}{35}[7P_0(x) + 20P_2(x) + 8P_4(x)] \end{aligned}$$

$$f(x) = \frac{1}{35}\,[7P_0(x) + 21P_1(x) + 20P_2(x) + 14P_3(x) + 8P_4(x)]$$

$$p_2^* = a_0^* P_0 + a_1^* P_1 + a_2^* P_2, \qquad a_k^* = \frac{\langle f, P_k \rangle}{\langle P_k, P_k \rangle}$$

$$a_0^* = \frac{7}{35}, \qquad a_1^* = \frac{21}{35}, \qquad a_2^* = \frac{20}{35}$$

$$p_2^*(x) = \frac{3}{35}(-1 + 7x + 10x^2)$$

(b)

$$T_0(x) = 1 \qquad 1 = T_0(x)$$

$$T_1(x) = x \qquad x = T_1(x)$$

$$T_2(x) = 2x^2 - 1 \qquad x^2 = \frac{1}{2}[T_0(x) + T_2(x)]$$

$$T_3(x) = 4x^3 - 3x \qquad x^3 = \frac{1}{4}[3T_1(x) + T_3(x)]$$

$$T_4(x) = 8x^4 - 8x^2 + 1 \qquad x^4 = \frac{1}{8}[3T_0(x) + 4T_2(x) + T_4(x)]$$

$$f(x) = \frac{1}{8}\,[3T_0(x) + 6T_1(x) + 4T_2(x) + 2T_3(x) + T_4(x)]$$

$$p_2^* = a_0^* T_0 + a_1^* T_1 + a_2^* T_2, \qquad a_k^* = \frac{\langle f, T_k\rangle}{\langle T_k, T_k\rangle}$$

$$a_0^* = \frac{3}{8}, \qquad a_1^* = \frac{3}{4}, \qquad a_2^* = \frac{1}{2}$$

$$p_2^*(x) = \frac{1}{8}\,(-1 + 6x + 8x^2)$$

8. INTEGRAÇÃO NUMÉRICA

Introdução.

• Neste capítulo derivamos e analisamos métodos numéricos para calcular integrais definidos da forma

$$I(f) = \int_a^b f(x)\,dx,$$

com $[a, b]$ finito. Estes métodos são necessários para o cálculo de integrais tais que:

- ◇ a primitiva da função integranda não é conhecida;
- ◇ embora conhecida a primitiva da função integranda é demasiado complicada, tornando mais rápido o cálculo numérico do integral;
- ◇ a integranda é conhecida apenas num número finito de pontos.

Além disso estes métodos de *integração numérica* ou *quadratura numérica* constituem uma ferramenta básica para a resolução numérica de equações diferenciais e equações integrais.

A maior parte dos métodos de integração numérica para calcular $I(f)$ encaixa-se no seguinte esquema: sendo f_n uma função aproximadora de f, define-se a **fórmula de integração ou quadratura numérica** por

$$I_n(f) := I(f_n).$$

f_n deve ser tal que $I(f_n)$ possa ser calculado facilmente. O erro de integração será calculado a partir do erro da função aproximadora f_n em relação a f:

$$E_n(f) = I(f) - I_n(f) = I(f - f_n).$$

A maior parte das fórmulas de integração numérica usam para funções aproximadoras f_n os polinómios interpoladores p_n ou funções interpoladoras seccionalmente polinomiais. Estudaremos seguidamente:

- ◇ as fórmulas de integração de Newton-Cotes, em que os polinómios interpoladores são suportados em nós igualmente espaçados;
- ◇ as fórmulas de integração de Gauss, em que os polinómios interpoladores são suportados em nós cuidadosamente escolhidos, e que não são igualmente espaçados; estas fórmulas são óptimas num certo sentido e têm convergência muito rápida.
- ◇ as fórmulas compostas correspondentes.

As fórmulas de integração numérica assim obtidas têm a forma

$$I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}), \qquad n \in \mathbb{N}.$$

Os coeficientes $w_{j,n}$ são chamados os **pesos de integração** ou **pesos de quadratura**; os pontos $x_{j,n}$ são chamados os **nós de integração** ou **nós de quadratura**, normalmente

escolhidos em $[a, b]$. A dependência em n é em geral suprimida, escrevendo-se w_j e x_j, mas subentendida implicitamente. Os métodos usuais de integração numérica têm nós e pesos com forma simples ou que são fornecidos em tabelas facilmente acessíveis. Não há em geral necessidade de construir explicitamente as funções aproximadoras f_n, embora seja útil ter presente o seu papel em definir $I_n(f)$.

Fórmulas de quadratura interpolatória polinomial (FQIP)

Proposição. Seja $f \in C([a, b])$ e $p_n \in \mathcal{P}_n$ o polinómio interpolador de f nos pontos $x_0, x_1, \ldots, x_n$ distintos do intervalo $[a, b]$. A FQIP de ordem n de f é definida por:

$$I_n(f) := I(p_n),$$

e tem a forma

$$I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}),$$

com pesos

$$w_{j,n} = I(l_{j,n}) = \int_a^b \prod_{\substack{i=0 \\ i \neq j}} \frac{x - x_{i,n}}{x_{j,n} - x_{i,n}}\, dx.$$

Dem.: $(\cdots)$

Notas. (i) $\mathcal{P}_n$ designa o conjunto de polinómios de grau menor ou igual a n.

(ii) $\displaystyle\sum_{j=0}^{n} w_{j,n} = b - a$

(iii) Sendo $f \in \mathcal{P}_n$, f coincide com o seu polinómio interpolador $p_n \in \mathcal{P}_n$ e, portanto,

$$I_n(f) = I(p_n) = I(f).$$

Proposição. Dados $n + 1$ pontos distintos $x_0, x_1, \ldots, x_n$ do intervalo $[a, b]$ a FQIP de ordem n é unicamente determinada pela propriedade de que integra exactamente todos os polinómios de grau menor ou igual a n, isto é,

$$I_n(q) = I(q), \qquad \forall q \in \mathcal{P}_n.$$

Dem.: $(\cdots)$

Nota. Esta condição traduz-se no sistema de equações

$$I_n(x^k) = I(x^k), \qquad k = 0, 1, \ldots, n.$$

Trata-se de um sistema de equações lineares nos pesos de integração. Este método de determinar as FQIP é muitas vezes designado por *método dos coeficientes indeterminados*.

Exemplo. Determinar as FQIP de ordens 0, 1 e 2 pelas duas vias indicadas nas duas proposições, isto é, integração dos polinómios de Lagrange e método dos coeficientes indeterminados. Obtém-se:

(a) $n = 0, \;\; a \leq x_0 \leq b$

$$I_0(f) = w_0 f(x_0) = (b-a) f(x_0)$$

Caso particular: $x_0 = \dfrac{a+b}{2}$ (Fórmula do ponto médio)

(b) $n = 1, \;\; a \leq x_0 < x_1 \leq b$

$$I_1(f) = w_0 f(x_0) + w_1 f(x_1)$$

$$w_0 = \frac{(b-a)(a+b-2x_1)}{2(x_0-x_1)}, \qquad w_1 = \frac{(b-a)(a+b-2x_0)}{2(x_1-x_0)}$$

Casos particulares:

(i) $x_0 = a, \quad x_1 = b, \qquad w_0 = \dfrac{b-a}{2} = w_1$

(Fórmula do trapézio ou fórmula de Newton-Cotes fechada de ordem 1)

(ii) $x_0 = a + \dfrac{b-a}{3}, \quad x_1 = b - \dfrac{b-a}{3}, \qquad w_0 = \dfrac{b-a}{2} = w_1$

(Fórmula de Newton-Cotes aberta de ordem 1)

(iii) $x_0 = \dfrac{a+b}{2} - \dfrac{b-a}{2}\,\dfrac{\sqrt{3}}{3}, \quad x_1 = \dfrac{a+b}{2} + \dfrac{b-a}{2}\,\dfrac{\sqrt{3}}{3}, \qquad w_0 = \dfrac{b-a}{2} = w_1$

(Fórmula de Gauss-Legendre de ordem 1)

(c) $n = 2, \;\; a \leq x_0 < x_1 < x_2 \leq b$

$$I_2(f) = w_0 f(x_0) + w_1 f(x_1) + w_2 f(x_2)$$

$$w_0 = \frac{(b-a)[2(a^2+b^2+ab) + 6x_1x_2 - 3(a+b)(x_1+x_2)]}{6(x_0-x_1)(x_0-x_2)}$$

$$w_1 = \frac{(b-a)[2(a^2+b^2+ab) + 6x_2x_0 - 3(a+b)(x_2+x_0)]}{6(x_1-x_2)(x_1-x_0)}$$

$$w_2 = \frac{(b-a)[2(a^2+b^2+ab) + 6x_0x_1 - 3(a+b)(x_0+x_1)]}{6(x_2-x_0)(x_2-x_1)}$$

Casos particulares:

(i) $x_0 = a, \quad x_1 = \dfrac{a+b}{2}, \quad x_2 = b, \qquad w_0 = \dfrac{b-a}{6} = w_2, \quad w_1 = 4\,\dfrac{b-a}{6}$

(Fórmula de Simpson ou fórmula de Newton-Cotes fechada de ordem 2)

(ii) $x_0 = a + \dfrac{b-a}{4}, \quad x_1 = \dfrac{a+b}{2}, \quad x_2 = b - \dfrac{b-a}{4}$

$w_0 = 2\,\dfrac{b-a}{3} = w_2, \quad w_1 = -\dfrac{b-a}{3}$

(Fórmula de Newton-Cotes aberta de ordem 2)

(iii) $x_0 = \dfrac{a+b}{2} - \dfrac{b-a}{2}\sqrt{\dfrac{3}{5}}, \quad x_1 = 0, \quad x_2 = \dfrac{a+b}{2} + \dfrac{b-a}{2}\sqrt{\dfrac{3}{5}}$

$w_0 = \dfrac{5}{9}\,\dfrac{b-a}{2} = w_2, \quad w_1 = \dfrac{8}{9}\,\dfrac{b-a}{2}$

(Fórmula de Gauss-Legendre de ordem 2)

Resolução: (b)

Método dos polinómios de Lagrange:

$$w_0 = \int_a^b l_0(x)\,dx = \int_a^b \frac{x-x_1}{x_0-x_1}\,dx = \frac{1}{x_0-x_1}\left[\frac{(x-x_1)^2}{2}\right]_a^b = \frac{(b-a)(a+b-2x_1)}{2(x_0-x_1)}$$

$$w_1 = \int_a^b l_1(x)\,dx = \int_a^b \frac{x-x_0}{x_1-x_0}\,dx = \frac{1}{x_1-x_0}\left[\frac{(x-x_0)^2}{2}\right]_a^b = \frac{(b-a)(a+b-2x_0)}{2(x_1-x_0)}$$

Método dos coeficientes indeterminados:

$$\begin{cases} I_1(1) = I(1) \\ I_1(x) = I(x) \end{cases} \Rightarrow \begin{cases} w_0 + w_1 = b-a \\ x_0 w_0 + x_1 w_1 = \dfrac{b^2-a^2}{2} \end{cases} \Rightarrow \begin{cases} w_0 = \dfrac{(b-a)(a+b-2x_1)}{2(x_0-x_1)} \\ w_1 = \dfrac{(b-a)(a+b-2x_0)}{2(x_1-x_0)} \end{cases}$$

Fórmulas de Newton-Cotes fechadas e abertas

Proposição. A FQIP de ordem $n \in \mathbb{N}_1$ com nós de integração equidistantes

$$x_{j,n} = a + jh_n, \qquad j = 0, 1, \ldots, n,$$

onde

$$h_n = \frac{b-a}{n},$$

é chamada a **fórmula de Newton-Cotes fechada de ordem** n. Os seus pesos de integração são dados por

$$w_{j,n} = h_n\,\frac{(-1)^{n-j}}{j!(n-j)!}\int_0^n \prod_{\substack{i=0 \\ i\neq j}}^{n} (t-i)\,dt,$$

e têm a simetria,

$$w_{j,n} = w_{n-j,n}, \qquad j = 0, 1, \ldots, n.$$

Exemplo. $I_1(f)$ e $I_2(f)$ foram calculados no exemplo anterior. $I_3(f), I_4(f), I_5(f), I_6(f), I_7(f)$ estão disponíveis no Anexo.

Proposição (Teorema do Valor Médio para Integrais). Sejam $f, u \in C([a,b]),\ u(x) \geq 0,\ \forall x \in [a,b]$. Então:

$$\int_a^b u(x) f(x)\, dx = f(\xi) \int_a^b u(x)\, dx,$$

para algum $\xi \in [a,b]$.

Proposição. Seja $f \in C^2([a,b])$. O erro de integração para a fórmula do trapézio é dado por

$$E_1(f) = I(f) - I_1(f) = -\frac{h^3}{12} f''(\xi),$$

onde $h = b - a$ e $\xi \in]a,b[$.

Dem.: $(\cdots)$

Proposição. Seja $f \in C^4([a,b])$. O erro de integração para a fórmula de Simpson é dado por

$$E_2(f) = I(f) - I_2(f) = -\frac{h^5}{90} f^{(4)}(\xi),$$

onde $h = \dfrac{b-a}{2}$ e $\xi \in]a,b[$.

Dem.: $(\cdots)$

Proposição. Seja $f \in C^{n+\nu_n}([a,b])$, onde $\nu_n = 1 + \dfrac{1}{2}\,[1 + (-1)^n]$; isto é, $\nu_n = 1$ para n ímpar e $\nu_n = 2$ para n par. Então o erro de integração para a fórmula de Newton-Cotes fechada de ordem n é dado por

$$E_n(f) = I(f) - I_n(f) = \frac{C_n}{(n+\nu_n)!}\, h_n^{n+1+\nu_n} f^{(n+\nu_n)}(\xi),$$

onde

$$C_n = \int_0^n t^{\nu_n - 1} \prod_{i=0}^{n} (t - i)\, dt,$$

$h_n = \dfrac{b-a}{n}$ e $\xi \in]a,b[$.

Nota. Mostra-se que $C_n < 0,\ \forall n \in \mathbb{N}_1$.

Exemplo. $E_3(f), \ldots, E_7(f)$ estão disponíveis no Anexo.

Definição. Uma fórmula de integração numérica $I_n(f)$ que aproxima $I(f)$ diz-se de **grau**

de precisão m se:

(1) $I_n(q) = I(q), \quad \forall q \in \mathcal{P}_m;$

(2) $I_n(q) \neq I(q), \quad$ para algum $q \in \mathcal{P}_{m+1}$.

Proposição. As fórmulas de Newton-Cotes fechadas de ordem n têm grau de precisão $n + \nu_n - 1$, isto é, têm grau de precisão

(1) n, para n ímpar; (2) $n+1$, para n par.

Proposição. A FQIP de ordem $n \in \mathbb{N}$ com nós de integração equidistantes

$$x_{j,n} = a + (j+1)h_n, \qquad j = 0, 1, \ldots, n,$$

onde

$$h_n = \frac{b-a}{n+2},$$

é chamada a **fórmula de Newton-Cotes aberta de ordem** n. Os seus pesos de integração são dados por

$$w_{j,n} = h_n \frac{(-1)^{n-j}}{j!(n-j)!} \int_{-1}^{n+1} \prod_{\substack{i=0 \\ i \neq j}}^{n} (t-i)\, dt,$$

e têm a simetria,

$$w_{j,n} = w_{n-j,n}, \qquad j = 0, 1, \ldots, n.$$

Exemplo. $I_0(f), I_1(f), I_2(f)$ foram calculadas no exemplo de FQIP.

Proposição. Seja $f \in C^2([a,b])$. O erro de integração para a fórmula do ponto médio é dado por

$$E_0(f) = I(f) - I_0(f) = \frac{h^3}{3} f''(\xi),$$

onde $h = \dfrac{b-a}{2}$ e $\xi \in]a,b[$.

Dem.: $(\cdots)$

Proposição. Seja $f \in C^{n+\nu_n}([a,b])$, onde $\nu_n = 1 + \dfrac{1}{2}\,[1 + (-1)^n]$. Então o erro de integração para a fórmula de Newton-Cotes aberta de ordem n é dado por

$$E_n(f) = I(f) - I_n(f) = \frac{C_n}{(n+\nu_n)!}\, h_n^{n+1+\nu_n} f^{(n+\nu_n)}(\xi),$$

onde

$$C_n = \int_{-1}^{n+1} t^{\nu_n - 1} \prod_{i=0}^{n} (t-i)\, dt,$$

$h_n = \dfrac{b-a}{n+2}$ e $\xi \in [a,b]$.

Nota. Mostra-se que $C_n > 0, \ \forall n \in \mathbb{N}$.

Exemplo. $E_1(f)$ e $E_2(f)$ estão disponíveis no Anexo.

Proposição. As fórmulas de Newton-Cotes abertas de ordem n têm grau de precisão $n + \nu_n - 1$, isto é, têm grau de precisão

$$(1) \ \ n, \ \text{para } n \text{ ímpar}; \qquad (2) \ \ n+1, \ \text{para } n \text{ par}.$$

Proposição. Seja $f \in C^{n+\nu_n}([a,b])$, onde $\nu_n = 1 + \frac{1}{2}\,[1 + (-1)^n]$. Então o erro de integração para a fórmula de Newton-Cotes, fechada ou aberta, de ordem n é dado por

$$E_n(f) = I(f) - I_n(f) = \frac{E_n(q)}{(n+\nu_n)!}\, f^{(n+\nu_n)}(\xi),$$

onde $q(x) = (x-a)^{n+\nu_n}$ e $\xi \in]a,b[$.

Exemplo. Determinar as expressões dos erros da fórmula dos trapézios e da fórmula de Simpson.

$$E_1(f) = \frac{E_1(q)}{2!}\, f''(\xi), \qquad q(x) = (x-a)^2$$

$$\begin{aligned} E_1(q) &= I(q) - I_1(q) = \int_a^b q(x)\,dx - \frac{b-a}{2}\,[q(a)+q(b)] \\ &= \frac{(b-a)^3}{3} - \frac{(b-a)^3}{2} = -\frac{(b-a)^3}{6} \end{aligned}$$

$$E_1(f) = -\frac{(b-a)^3}{12}\, f''(\xi)$$

$$E_2(f) = \frac{E_2(q)}{4!}\, f^{(4)}(\xi), \qquad q(x) = (x-a)^4$$

$$\begin{aligned} E_2(q) &= I(q) - I_2(q) = \int_a^b q(x)\,dx - \frac{b-a}{6}\left[q(a) + 4q\left(\frac{a+b}{2}\right) + q(b)\right] \\ &= \frac{(b-a)^5}{5} - \frac{5}{24}\,(b-a)^5 = -\frac{(b-a)^5}{120} \end{aligned}$$

$$E_2(f) = -\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi)$$

Fórmulas de Newton-Cotes fechadas e abertas compostas

• Uma vez que os erros das fórmulas de Newton-Cotes são proporcionais a potências de $b-a$, se esta quantidade não for suficientemente pequena, as fórmulas deixam de ter utilidade. Neste caso o que se deve fazer é dividir o intervalo $[a,b]$ em subintervalos e aplicar a cada um dos integrais assim obtidos uma das fórmulas de Newton-Cotes.

Proposição. Seja $I_n(f;[a,b])$ a fórmula de Newton-Cotes de ordem n, fechada ou aberta, para obter um valor aproximado do integral

$$I(f;[a,b]) = \int_a^b f(x)\,dx,$$

anteriormente designadas simplesmente por $I_n(f)$ e $I(f)$, respectivamente. Sejam $x_0, x_1, \ldots, x_M$ os pontos do intervalo $[a,b]$ definidos por

$$x_j = a + jh_M, \quad j = 0, 1, \ldots, M, \qquad h_M = \frac{b-a}{M},$$

onde $M \in \mathbb{N}_1$ é um múltiplo de $n+\mu$, onde $\mu = 0$ no caso das fórmulas fechadas, e $\mu = 2$, no caso das fórmulas abertas. Então a fórmula de Newton-Cotes, fechada ou aberta, de ordem n, composta, com M subintervalos, para obter um valor aproximado do integral $I(f)$, é

$$I_n^{(M)}(f) = \sum_{j=1}^{M/(n+\mu)} I_n(f;[x_{(n+\mu)(j-1)}, x_{(n+\mu)j}]).$$

Exemplo. Fórmula do trapézio composta

$$I_1^{(M)}(f) = \frac{h_M}{2}\left[f(x_0) + 2\sum_{j=1}^{M-1} f(x_j) + f(x_M)\right], \qquad h_M = \frac{b-a}{M}$$

Com efeito:

$$\begin{aligned} I_1^{(M)}(f) &= \sum_{j=1}^{M} I_1\left(f;[x_{j-1}, x_j]\right) \\ &= \sum_{j=1}^{M} \frac{1}{2}\left(x_j - x_{j-1}\right)\left[f(x_{j-1}) + f(x_j)\right] \\ &= \frac{b-a}{2M}\sum_{j=1}^{M}\left[f(x_{j-1}) + f(x_j)\right] \end{aligned}$$

Exemplo. Fórmula de Simpson composta

$$I_2^{(M)}(f) = \frac{h_M}{3}\left[f(x_0) + 2\sum_{j=1}^{M/2-1} f(x_{2j}) + 4\sum_{j=1}^{M/2} f(x_{2j-1}) + f(x_M)\right], \qquad h_M = \frac{b-a}{M}$$

Com efeito:

$$\begin{aligned} I_2^{(M)}(f) &= \sum_{j=1}^{M/2} I_2\left(f;[x_{2j-2}, x_{2j}]\right) \\ &= \sum_{j=1}^{M/2} \frac{1}{6}\left(x_{2j} - x_{2j-2}\right)\left[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\right] \\ &= \frac{b-a}{3M}\sum_{j=1}^{M/2}\left[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\right] \end{aligned}$$

Exemplo. As fórmulas de Newton-Cotes fechadas compostas para $n = 3, \ldots, 7$ e as fórmulas de Newton-Cotes abertas compostas para $n = 0, 1, 2$ estão disponíveis no Anexo.

Proposição. Seja $f \in C^2([a, b])$. O erro de integração para a fórmula do trapézio composta é dado por

$$E_1^{(M)}(f) = I(f) - I_1^{(M)}(f) = -\frac{b-a}{12}\, h_M^2 f''(\xi),$$

onde $h_M = \dfrac{b-a}{M}$ e $\xi \in]a, b[$.

Dem.: $(\cdots)$

Proposição. Seja $f \in C^{n+\nu_n}([a, b])$ onde $\nu_n = 1 + \dfrac{1}{2}\,[1 + (-1)^n]$. Seja $M \in \mathbb{N}_1$ um múltiplo de $n + \mu$, onde $\mu = 0$ no caso das fórmulas fechadas, e $\mu = 2$ no caso das fórmulas abertas. O erro de integração para a fórmula de Newton-Cotes de ordem n, fechada ou aberta, composta, com M subintervalos de integração, é dado por

$$E_n^{(M)}(f) = I(f) - I_n^{(M)}(f) = \frac{b-a}{n+\mu}\, \frac{C_n^\mu}{(n+\nu_n)!}\, h_M^{n+\nu_n}\, f^{(n+\nu_n)}(\xi),$$

onde $h_M = \dfrac{b-a}{M}$ e $\xi \in]a, b[$. Os coeficientes C_n^μ foram obtidos anteriormente em ambos os casos.

Exemplo. Os erros de integração para as fórmulas de Newton-Cotes fechadas compostas de ordens $2, \ldots, 7$ e abertas compostas de ordens $0, 1, 2$ estão disponíveis no Anexo.

Fórmulas de integração de Gauss

• Dados os nós de integração $x_{0,n}, x_{1,n}, \ldots, x_{n,n}$ em $[a, b]$, os pesos de integração $w_{0,n}, w_{1,n}, \ldots, w_{n,n}$ das FQIP

$$I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}),$$

são determinados por forma a que todos os polinómios de grau $\leq n$ sejam integrados exactamente, isto é,

$$I_n(p) = I(p), \qquad \forall p \in \mathcal{P}_n.$$

Vamos agora considerar o problema de escolher os nós $x_{0,n}, x_{1,n}, \ldots, x_{n,n}$ por forma a que a FQIP integre exactamente os polinómios de maior grau $m \geq n$ possível, isto é,

$$I_n(p) = I(p), \qquad \forall p \in \mathcal{P}_m,$$

e determinar esse grau.

Os nós e os pesos de quadratura, num total de $2n + 2$ incógnitas, devem satisfazer ao sistema de equações não-lineares

$$I_n(x^k) = I(x^k), \qquad k = 0, 1, \ldots, m.$$

Veremos que é efectivamente para $m = 2n + 1$, valor para o qual o número de equações coincide com o número de incógnitas, que este sistema tem uma única solução e que para esta solução os pontos $x_{0,n}, x_{1,n}, \ldots, x_{n,n}$ são distintos. São os zeros do polinómio de grau $n + 1$ de um sistema de polinómios ortogonais com respeito ao produto interno definido por

$$\langle f, g\rangle = I(fg), \qquad \forall f, g \in C([a, b]).$$

Para ganharmos alguma sensibilidade para a dificuldade do problema consideremos o seguinte exemplo:

Exemplo. No caso do integral

$$I(f) = \int_{-1}^{1} f(x)\, dx,$$

determinar as FQIP que integram exactamente todos os polinómios de grau menor ou igual a $2n + 1$, para $n = 0, 1, 2$.

O sistema de $2n + 2$ equações não lineares a resolver é:

$$\sum_{j=0}^{n} w_{j,n} x_{j,n}^{k} = \frac{1 - (-1)^{k+1}}{k + 1}, \qquad k = 0, 1, \ldots, 2n + 1,$$

isto é,

$$\sum_{j=0}^{n} w_{j,n} x_{j,n}^{k} = \begin{cases} 0, & k = 1, 3, \ldots, 2n + 1 \\ \dfrac{2}{k + 1}, & k = 0, 2, \ldots, 2n \end{cases}.$$

$$n = 0 : \begin{cases} w_{0,0} x_{0,0} = 0 \\ w_{0,0} = 2 \end{cases} \qquad \begin{cases} x_{0,0} = 0 \\ w_{0,0} = 2 \end{cases}$$

$$I_0(f) = 2f(0)$$

$$n = 1 : \begin{cases} w_{0,1} x_{0,1} + w_{1,1} x_{1,1} = 0 \\ w_{0,1} x_{0,1}^3 + w_{1,1} x_{1,1}^3 = 0 \\ w_{0,1} + w_{1,1} = 2 \\ w_{0,1} x_{0,1}^2 + w_{1,1} x_{1,1}^2 = \dfrac{2}{3} \end{cases} \qquad \begin{cases} x_{0,1} = -\dfrac{\sqrt{3}}{3} \\ x_{1,1} = \dfrac{\sqrt{3}}{3} \\ w_{0,1} = 1 \\ w_{1,1} = 1 \end{cases}$$

$$I_1(f) = f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right)$$

$$n=2:\begin{cases} w_{0,2}x_{0,2}+w_{1,2}x_{1,2}+w_{2,2}x_{2,2}=0 \\ w_{0,2}x_{0,2}^3+w_{1,2}x_{1,2}^3+w_{2,2}x_{2,2}^3=0 \\ w_{0,2}x_{0,2}^5+w_{1,2}x_{1,2}^5+w_{2,2}x_{2,2}^5=0 \\ w_{0,2}+w_{1,2}+w_{2,2}=2 \\ w_{0,2}x_{0,2}^2+w_{1,2}x_{1,2}^2+w_{2,2}x_{2,2}^2=\dfrac{2}{3} \\ w_{0,2}x_{0,2}^4+w_{1,2}x_{1,2}^4+w_{2,2}x_{2,2}^4=\dfrac{2}{5} \end{cases} \qquad \begin{cases} x_{0,2}=-\sqrt{\dfrac{3}{5}} \\ x_{1,2}=0 \\ x_{2,2}=\sqrt{\dfrac{3}{5}} \\ w_{0,2}=\dfrac{5}{9} \\ w_{1,2}=\dfrac{8}{9} \\ w_{2,2}=\dfrac{5}{9} \end{cases}$$

$$I_2(f)=\frac{5}{9}\,f\left(-\sqrt{\frac{3}{5}}\right)+\frac{8}{9}\,f(0)+\frac{5}{9}\,f\left(\sqrt{\frac{3}{5}}\right)$$

• A solução do sistema torna-se rapidamente demasiado complicada. A solução do problema passa pois por uma outra via.

• No estudo das fórmulas de quadratura de Gauss vamos considerar com mais generalidade o integral

$$I(f)=\int_a^b w(x)f(x)\,dx,$$

onde w é uma *função de peso*, anteriormente definida.

Definição. Uma FQIP

$$I_n(f)=\sum_{j=0}^{n} w_{j,n}f(x_{j,n}),$$

com $n+1$ nós de quadratura distintos $x_{0,n},\ldots,x_{n,n}$ é chamada uma **fórmula de quadratura de Gauss** se integra exactamente todos os polinómios de grau menor ou igual a $2n+1$, isto é,

$$I_n(p)=I(p), \qquad \forall p\in\mathcal{P}_{2n+1}.$$

Proposição. Para cada $n\in\mathbb{N}$ existe uma única fórmula de quadratura de Gauss. Os seus nós de quadratura são os zeros do polinómio de grau $n+1$ pertencente ao sistema de polinómios ortogonais $\{\varphi_k\}$ com respeito ao produto interno

$$\langle f,g\rangle=I(fg), \qquad \forall f,g\in C([a,b]).$$

Os seus pesos de quadratura são determinados pelas fórmulas

$$w_{j,n}=I(l_{j,n}), \qquad j=0,1,\ldots,n,$$

onde $l_{0,n},\ldots,l_{n,n}$ são os polinómios de Lagrange de grau n suportados nos nós de inte-

gração, ou, por qualquer dos sistemas (lineares) de $n+1$ equações:

$$\text{(a)} \quad \sum_{j=0}^{n} w_{j,n} x_{j,n}^k = I(x^k), \qquad k = 0, 1, \ldots, n;$$

$$\text{(b)} \quad \sum_{j=0}^{n} w_{j,n} \varphi_k(x_{j,n}) = I(\varphi_k), \qquad k = 0, 1, \ldots, n.$$

Proposição. A fórmula de quadratura de Gauss de ordem n tem grau de precisão $2n+1$.

Dem.: $(\cdots)$

Proposição. Os pesos da fórmula de quadratura de Gauss de ordem n são dados por

$$w_{j,n} = I(l_{j,n}^2), \qquad j = 0, 1, \ldots, n.$$

Em particular

$$w_{j,n} > 0, \qquad j = 0, 1, \ldots, n.$$

Dem.: $(\cdots)$

Proposição. Os pesos da fórmula de quadratura de Gauss de ordem n são dados por

$$w_{j,n} = -\frac{I(\Phi_{n+1}^2)}{\Phi'_{n+1}(x_{j,n})\Phi_{n+2}(x_{j,n})}, \qquad j = 0, 1, \ldots, n$$

onde Φ_n é o elemento de grau n do sistema de polinómios mónicos ortogonais em relação ao produto interno

$$\langle f, g \rangle = I(fg), \qquad \forall f, g \in C([a, b]).$$

Proposição. Seja $f \in C^{2n+2}([a, b])$. Então o erro da fórmula de quadratura de Gauss de ordem n é dado por

$$E_n(f) = I(f) - I_n(f) = \frac{I(\Phi_{n+1}^2)}{(2n+2)!} f^{(2n+2)}(\xi),$$

para algum $\xi \in]a, b[$.

Dem.: $(\cdots)$

Exemplo. **Fórmulas de Gauss-Legendre** $([a, b] = [-1, 1], \;\; w(x) \equiv 1)$

$$I(f) = \int_{-1}^{1} f(x)\, dx \;\; \approx \;\; I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n})$$

$x_{j,n}$, $j = 0, 1, \ldots, n$: zeros do polinómio de Legendre P_{n+1}

$$w_{j,n} = -\frac{2}{(n+2)P'_{n+1}(x_{j,n})P_{n+2}(x_{j,n})}, \qquad j = 0, 1, \ldots, n$$

$$E_n(f) = \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\, f^{(2n+2)}(\xi), \qquad f \in C^{2n+2}([-1,1]), \quad \xi \in]a,b[$$

Estes resultados obtêm-se combinando nas expressões gerais para $w_{j,n}$ e $E_n(f)$ com as seguintes propriedades dos polinómios de Legendre (Capítulo 7):

$$\Phi_n = \frac{P_n}{A_n}, \qquad A_n = \frac{(2n)!}{2^n(n!)^2}, \qquad \langle P_n, P_n\rangle = \frac{2}{2n+1}.$$

As fórmulas de Gauss-Legendre de ordens 0, 1, 2 foram obtidas anteriormente. A fórmula de ordem 3 está disponível no Anexo. Também os erros destas quatro fórmulas estão disponíveis no Anexo.

Proposição. Seja $f \in C^{2n+2}([a,b])$ uma função tal que

$$\sup_{k\geq 0} M_k < \infty, \qquad M_k := \max_{x\in[-1,1]} \frac{\left|f^{(k)}(x)\right|}{k!}.$$

Então o erro da fórmula de quadratura de Gauss-Legendre de ordem n satisfaz a

$$|E_n(f)| \leq \frac{\pi}{4^{n+1}}\, M_{2n+2}, \qquad n \to \infty.$$

Nota. Este resultado mostra que $E_n(f) \to 0$, $n \to \infty$, com um decrescimento exponencial. Compare-se com o decrescimento da forma $1/M^{n+\nu_n}$, $M \to \infty$, para as fórmulas de Newton-Cotes compostas de ordem n.

Exemplo. **Fórmulas de Gauss-Chebyshev** $\left([a,b] = [-1,1],\ \ w(x) = 1/\sqrt{1-x^2}\right)$:

$$I(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx \quad \approx \quad I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n})$$

$$x_{j,n} = -\cos\left(\frac{2j+1}{2n+2}\,\pi\right), \qquad w_{j,n} = \frac{\pi}{n+1}, \qquad j = 0, 1, \ldots, n$$

$$E_n(f) = \frac{\pi}{2^{2n+1}(2n+2)!}\, f^{(2n+2)}(\xi), \qquad \xi \in]a,b[$$

Estes resultados obtêm-se combinando nas expressões gerais para $w_{j,n}$ e $E_n(f)$ com as seguintes propriedades dos polinómios de Chebyshev (Capítulo 7):

$$\Phi_n = \frac{T_n}{2^{n-1}}, \qquad \langle T_n, T_n\rangle = \frac{\pi}{2}, \qquad T_n(x) = \cos(n \arccos x)$$

No caso dos pesos de integração temos:

$$w_{j,n} = -\frac{I(\Phi_{n+1}^2)}{\Phi'_{n+1}(x_{j,n})\Phi_{n+2}(x_{j,n})} = -\frac{\pi}{T'_{n+1}(x_{j,n})T_{n+2}(x_{j,n})}$$

$$x = \cos\theta, \qquad T_n(x) = \cos(n\theta), \qquad T'_n(x) = \frac{n\sin(n\theta)}{\sin\theta}$$

$$x_{j,n} = \cos\theta_{j,n}, \qquad \theta_{j,n} = \pi - \frac{\pi(2j+1)}{2(n+1)}$$

$$w_{j,n} = -\frac{\pi\sin\theta_{j,n}}{(n+1)\sin[(n+1)\theta_{j,n}]\cos[(n+2)\theta_{j,n}]} = \frac{\pi}{n+1}$$

Exemplo. Considere o integral

$$I(f) = \int_{-b}^{b} w(x)f(x)\,dx,$$

onde $f, w \in C([-b,b])$, w é definida por $w(x) = b - |x|$ e b é uma constante positiva. Determine as fórmulas de quadratura de Gauss de ordens 1, 2 e 3 para aproximar o integral $I(f)$. Note que

$$I(x^m) = \begin{cases} \dfrac{2b^{m+2}}{(m+2)(m+1)}, & m \text{ par} \\ 0, & m \text{ ímpar} \end{cases}$$

Resolução:

– Construção do conjunto de polinómios ortogonais em relação ao produto interno definido por

$$\langle f, g\rangle = I(fg), \qquad \forall f, g \in C([-b,b])$$

$$\varphi_0(x) = 1$$

$$\varphi_1(x) = x - B_1 = x$$

$$B_1 = \frac{\langle x\varphi_0, \varphi_0\rangle}{\langle \varphi_0, \varphi_0\rangle} = \frac{I(x)}{I(1)} = 0$$

$$\varphi_2(x) = (x - B_2)\varphi_1(x) - C_2\varphi_0(x) = x^2 - \frac{b^2}{6}$$

$$B_2 = \frac{\langle x\varphi_1, \varphi_1\rangle}{\langle \varphi_1, \varphi_1\rangle} = \frac{I(x^3)}{I(x^2)} = 0$$

$$C_2 = \frac{\langle x\varphi_1, \varphi_0\rangle}{\langle \varphi_0, \varphi_0\rangle} = \frac{I(x^2)}{I(1)} = \frac{b^2}{6}$$

$$\varphi_3(x) = (x - B_3)\varphi_2(x) - C_3\varphi_1(x) = x\left(x^2 - \frac{2b^2}{5}\right)$$

$$B_3 = \frac{\langle x\varphi_2, \varphi_2\rangle}{\langle \varphi_2, \varphi_2\rangle} = \frac{I(x[\varphi_2(x)]^2)}{I([\varphi_2(x)]^2)} = 0$$

$$C_3 = \frac{\langle x\varphi_2, \varphi_1\rangle}{\langle \varphi_1, \varphi_1\rangle} = \frac{I\left(x^4 - \frac{b^2}{6}x^2\right)}{I(x^2)} = \frac{7b^2}{30}$$

– Cálculo das fórmulas de Gauss:

$$n=0: \quad I_0(f)=w_0 f(x_0), \qquad x_0=0$$

$$w_0 = I(1) = b^2$$

$$n=1: \quad I_1(f)=w_0 f(x_0)+w_1 f(x_1), \qquad x_1=\frac{b}{\sqrt{6}}=-x_0$$

$$\begin{cases} w_0+w_1 = I(1) = b^2 \\ w_0 x_0 + w_1 x_1 = I(x) = 0 \end{cases} \Rightarrow \quad w_0 = w_1 = \frac{b^2}{2}$$

$$n=2: \quad I_2(f)=w_0 f(x_0)+w_1 f(x_1)+w_2 f(x_2), \qquad x_2 = b\sqrt{\frac{2}{5}} = -x_0, \;\; x_1 = 0$$

$$\begin{cases} w_0+w_1+w_2 = I(1) = b^2 \\ w_0 x_0 + w_1 x_1 + w_2 x_2 = I(x) = 0 \\ w_0 x_0^2 + w_1 x_1^2 + w_2 x_2^2 = I(x^2) = \dfrac{b^4}{6} \end{cases} \Rightarrow \begin{cases} w_0 = w_2 = \dfrac{5b^2}{24} \\ w_1 = \dfrac{7b^2}{12} \end{cases}$$

Fórmula de Gauss-Legendre composta

Proposição (da mudança de variável). Seja

$$I_n(f;[-1,1]) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n})$$

uma FQIP para obter um valor aproximado do integral

$$I(f;[-1,1]) = \int_{-1}^{1} f(t)\,dt.$$

Então a FQIP

$$I_n(f;[a,b]) = \sum_{j=0}^{n} w^*_{j,n} f(x^*_{j,n}),$$

onde

$$w^*_{j,n} = \frac{b-a}{2}\, w_{j,n}, \qquad x^*_{j,n} = a + \frac{b-a}{2}(x_{j,n}+1),$$

permite obter um valor aproximado para o integral

$$I(f;[a,b]) = \int_{a}^{b} f(x)\,dx.$$

Dem.: $(\cdots)$

Proposição. Seja

$$I_n(f;[-1,1]) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}),$$

a fórmula de Gauss-Legendre para obter um valor aproximado do integral

$$I(f;[-1,1]) = \int_{-1}^{1} f(t)\,dt.$$

Então a **fórmula de Gauss-Legendre composta** com M subintervalos para obter um valor aproximado do integral

$$I(f;[a,b]) = \int_{a}^{b} f(x)\,dx.$$

é dada por

$$I_n^{(M)}(f;[a,b]) = \frac{h_M}{2} \sum_{j=0}^{n} w_{j,n} \sum_{m=1}^{M} f\left(x_{j,n}^{(m)}\right),$$

onde

$$x_{j,n}^{(m)} = a + h_M(m-1) + \frac{h_M}{2}\left(x_{j,n}+1\right), \qquad h_M = \frac{b-a}{M}.$$

Dem.: $(\cdots)$

Proposição. Seja $f \in C^{2n+2}([a,b])$. O erro de integração da fórmula de Gauss-Legendre composta de ordem n com M subintervalos de integração é dado por

$$E_n^{(M)}(f) = \frac{b-a}{2}\left(\frac{h_M}{2}\right)^{2n+2} E_n(f),$$

onde

$$E_n(f) = \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\, f^{(2n+2)}(\xi), \qquad \xi \in]a,b[$$

Convergência de fórmulas de quadratura

Definição. Diz-se que uma sucessão de fórmulas de quadratura que aproxima o integral $I(f)$ é **convergente** se

$$I_n(f) \to I(f), \quad n \to \infty, \qquad \forall f \in C([a,b]).$$

Proposição. Seja

$$I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}),$$

uma sucessão de fórmulas de quadratura que aproxima o integral $I(f)$ tal que:

(1) $I_n(p) \to I(p), \quad n \to \infty,$ para qualquer polinómio p;
(2) $w_{j,n} > 0, \quad \forall j, \ \forall n.$

Então a sucessão é convergente.

Proposição.

(1) As fórmulas de quadratura de Newton-Cotes fechadas compostas de ordens 1 a 7 são convergentes, isto é,

$$I_n^{(M)}(f) \to I(f), \quad M \to \infty, \qquad \forall f \in C([a,b]).$$

(2) As fórmulas de quadratura de Gauss de ordem n são convergentes, isto é,

$$I_n(f) \to I(f), \quad n \to \infty, \qquad \forall f \in C([a,b]).$$

Nota. As fórmulas de quadratura de Newton-Cotes de ordem n não são convergentes, isto é,

$$I_n(f) \not\to I(f), \quad n \to \infty, \qquad f \in C([a,b]).$$

Anexo

- Fórmulas de Newton-Cotes fechadas de ordem n:

  - $n = 1,\ h = b - a$ (Regra dos trapézios):

$$I_1(f) = \frac{b-a}{2}\,[f(a) + f(b)], \qquad E_1(f) = -\frac{h^3}{12} f''(\xi), \quad \xi \in]a, b[$$

  - $n = 2,\ h = \dfrac{b-a}{2}$ (Regra de Simpson):

$$I_2(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right], \qquad E_2(f) = -\frac{h^5}{90} f^{(4)}(\xi)$$

  - $n = 3,\ h = \dfrac{b-a}{3}$ (Regra dos três oitavos):

$$I_3(f) = \frac{b-a}{8}\,[f(a) + 3f(a+h) + 3f(b-h) + f(b)], \quad E_3(f) = -\frac{3h^5}{80} f^{(4)}(\xi)$$

  - $n = 4,\ h = \dfrac{b-a}{4}$ (Regra de Milne):

$$I_4(f) = \frac{b-a}{90}\left[7f(a) + 32f(a+h) + 12f\left(\frac{a+b}{2}\right) + 32f(b-h) + 7f(b)\right]$$

$$E_4(f) = -\frac{8h^7}{945} f^{(6)}(\xi), \qquad \xi \in]a, b[$$

  - $n = 5,\ h = \dfrac{b-a}{5}$:

$$I_5(f) = \frac{b-a}{288}\,[19f(a) + 75f(a+h) + 50f(a+2h) + 50f(b-2h) + 75f(b-h) + 19f(b)]$$

$$E_5(f) = -\frac{275h^7}{12096} f^{(6)}(\xi), \qquad \xi \in]a, b[$$

  - $n = 6,\ h = \dfrac{b-a}{6}$ (Regra de Weddle):

$$I_6(f) = \frac{b-a}{840}\Bigg[41f(a) + 216f(a+h) + 27f(a+2h) + 272f\left(\frac{a+b}{2}\right) + 27f(b-2h) + 216f(b-h) + 41f(b)\Bigg]$$

$$E_6(f) = -\frac{9h^9}{1400} f^{(8)}(\xi), \qquad \xi \in]a, b[$$

  - $n = 7,\ h = \dfrac{b-a}{7}$:

$$I_7(f) = \frac{b-a}{120960}\Big[5257f(a) + 25039f(a+h) + 9261f(a+2h) + 20923f(a+3h) + 20923f(b-3h) + 9261f(b-2h) + 25039f(b-h) + 5257f(b)\Big]$$

$$E_7(f) = -\frac{8183h^9}{518400} f^{(8)}(\xi), \qquad \xi \in]a,b[$$

• Fórmulas de Newton-Cotes abertas de ordem n:

∘• $n = 0,\ h = \dfrac{b-a}{2}$ (Regra do ponto médio):

$$I_0(f) = (b-a)\, f\left(\frac{a+b}{2}\right), \qquad E_0(f) = \frac{h^3}{3} f''(\xi), \quad \xi \in]a,b[$$

∘• $n = 1,\ h = \dfrac{b-a}{3}$:

$$I_1(f) = \frac{b-a}{2}\,\left[f(a+h) + f(b-h)\right], \qquad E_1(f) = \frac{3h^3}{4} f''(\xi), \quad \xi \in]a,b[$$

∘• $n = 2,\ h = \dfrac{b-a}{4}$:

$$I_2(f) = \frac{b-a}{3}\,\left[2f(a+h) - f\left(\frac{a+b}{2}\right) + 2f(b-h)\right]$$

$$E_2(f) = \frac{14h^5}{45} f^{(4)}(\xi), \quad \xi \in]a,b[$$

• Fórmulas de Newton-Cotes fechadas compostas:

∘• $n = 1$:

$$I_1^{(M)}(f) = \frac{h_M}{2}\left[f_0 + f_M + 2\sum_{j=1}^{M-1} f_j\right]$$

$$E_1^{(M)}(f) = -\frac{b-a}{12}\, h_M^2 f''(\xi), \qquad \xi \in]a,b[$$

∘• $n = 2$ (M par):

$$I_2^{(M)}(f) = \frac{h_M}{3}\left[f_0 + f_M + 4\sum_{j=1}^{M/2} f_{2j-1} + 2\sum_{j=1}^{M/2-1} f_{2j}\right]$$

$$E_2^{(M)}(f) = -\frac{b-a}{180}\, h_M^4 f^{(4)}(\xi), \qquad \xi \in]a,b[$$

∘• $n = 3$ (M múltiplo de 3):

$$I_3^{(M)}(f) = \frac{3h_M}{8}\left[f_0 + f_M + 2\sum_{j=1}^{M/3-1} f_{3j} + 3\sum_{j=1}^{M/3} (f_{3j-1} + f_{3j-2})\right]$$

$$E_3^{(M)}(f) = -\frac{b-a}{80}\, h_M^4 f^{(4)}(\xi), \qquad \xi \in]a,b[$$

$\circ\bullet$ $n = 4$ (M múltiplo de 4):

$$I_4^{(M)}(f) = \frac{4h_M}{90}\left[7\,(f_0 + f_M) + 14\sum_{j=1}^{M/4-1} f_{4j} + 32\sum_{j=1}^{M/4}(f_{4j-1} + f_{4j-3}) + 12\sum_{j=1}^{M/4} f_{4j-2}\right]$$

$$E_4^{(M)}(f) = -\frac{2(b-a)}{945}\,h_M^6 f^{(6)}(\xi), \qquad \xi \in]a,b[$$

$\circ\bullet$ $n = 5$ (M múltiplo de 5):

$$I_5^{(M)}(f) = \frac{5h_M}{288}\left[19\,(f_0 + f_M) + 38\sum_{j=1}^{M/5-1} f_{5j} + 75\sum_{j=1}^{M/5}(f_{5j-1} + f_{5j-4}) + 50\sum_{j=1}^{M/5}(f_{5j-2} + f_{5j-3})\right]$$

$$E_5^{(M)}(f) = -\frac{55(b-a)}{12096}\,h_M^6 f^{(6)}(\xi), \qquad \xi \in]a,b[$$

$\circ\bullet$ $n = 6$ (M múltiplo de 6):

$$I_6^{(M)}(f) = \frac{h_M}{140}\left[41\,(f_0 + f_M) + 82\sum_{j=1}^{M/6-1} f_{6j} + 216\sum_{j=1}^{M/6}(f_{6j-1} + f_{6j-5}) + 27\sum_{j=1}^{M/6}(f_{6j-2} + f_{6j-4}) + 272\sum_{j=1}^{M/6} f_{6j-3}\right]$$

$$E_6^{(M)}(f) = -\frac{3(b-a)}{2800}\,h_M^8 f^{(8)}(\xi), \qquad \xi \in]a,b[$$

$\circ\bullet$ $n = 7$ (M múltiplo de 7):

$$I_7^{(M)}(f) = \frac{h_M}{17280}\left[5257\,(f_0 + f_M) + 10514\sum_{j=1}^{M/7-1} f_{7j} + 25039\sum_{j=1}^{M/7}(f_{7j-1} + f_{7j-6}) + 9261\sum_{j=1}^{M/7}(f_{7j-2} + f_{7j-5}) + 20923\sum_{j=1}^{M/7}(f_{7j-3} + f_{7j-4})\right]$$

$$E_7^{(M)}(f) = -\frac{1169(b-a)}{518400}\,h_M^8 f^{(8)}(\xi), \qquad \xi \in]a,b[$$

• Fórmulas de Newton-Cotes abertas compostas:

$\circ\bullet$ $n = 0$ (M par):

$$I_0^{(M)}(f) = 2h_M\sum_{j=1}^{M/2} f_{2j-1}, \qquad E_0^{(M)}(f) = \frac{(b-a)h_M^2}{6}f''(\xi), \quad \xi \in]a,b[$$

∘• $n = 1$ (M múltiplo de 3):

$$I_1^{(M)}(f) = \frac{3h_M}{2} \sum_{j=1}^{M/3} (f_{3j-2} + f_{3j-1})$$

$$E_1^{(M)}(f) = \frac{b-a}{4}\, h_M^2 f''(\xi), \qquad \xi \in]a, b[$$

∘• $n = 2$ (M múltiplo de 4):

$$I_2^{(M)}(f) = \frac{4h_M}{3} \sum_{j=1}^{M/4} (2f_{4j-3} - f_{4j-2} + 2f_{4j-1})$$

$$E_2^{(M)}(f) = \frac{7(b-a)}{90}\, h_M^4 f^{(4)}(\xi), \qquad \xi \in]a, b[$$

• Fórmulas de Gauss-Legendre:

∘• $n = 0$

$$I_0(f) = 2f(0), \qquad E_0(f) = \frac{1}{3}\, f''(\xi), \quad \xi \in]a, b[$$

∘• $n = 1$

$$I_1(f) = f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right), \qquad E_1(f) = \frac{1}{135}\, f^{(4)}(\xi), \quad \xi \in]a, b[$$

∘• $n = 2$

$$I_2(f) = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}\, f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

$$E_2(f) = \frac{1}{15750}\, f^{(6)}(\xi), \qquad \xi \in]a, b[$$

∘• $n = 3$

$$I_3(f) = w_{0,3} f(x_{0,3}) + w_{1,3} f(x_{1,3}) + w_{2,3} f(x_{2,3}) + w_{3,3} f(x_{3,3})$$

$$x_{0,3} = -\sqrt{\frac{1}{7}\left(3 + 2\sqrt{\frac{6}{5}}\right)} = -x_{3,3}, \quad x_{1,3} = -\sqrt{\frac{1}{7}\left(3 - 2\sqrt{\frac{6}{5}}\right)} = -x_{2,3}$$

$$w_{0,3} = \frac{1}{6}\left(3 - \sqrt{\frac{5}{6}}\right) = w_{3,3}, \quad w_{1,3} = \frac{1}{6}\left(3 + \sqrt{\frac{5}{6}}\right) = w_{2,3}$$

$$E_3(f) = \frac{1}{3472875}\, f^{(8)}(\xi), \qquad \xi \in]a, b[$$

10. RESOLUÇÃO NUMÉRICA DE EQUAÇÕES DIFERENCIAIS ORDINÁRIAS: PROBLEMAS DE VALOR INICIAL

Introdução: Problemas de Valor Inicial (PVI)

• As equações diferenciais são uma das mais importantes ferramentas matemáticas usadas na modelação de problemas em Física, Química e Engenharia. Neste capítulo derivamos e analisamos métodos numéricos para resolver problemas para equações diferenciais ordinárias (EDO's).

• Vamos considerar em primeiro lugar o chamado **problema de valor inicial** ou **problema de Cauchy**.

Definição. Seja $f : D \subset \mathbb{R}^{1+1} \to \mathbb{R}$ uma função contínua numa região D e $(x_0, Y_0) \in D$. O **PVI**

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(x_0) = Y_0, \end{cases} \tag{P}$$

é o problema de determinar uma função $Y : I_\alpha \to \mathbb{R}, \;\; I_\alpha = [x_0 - \alpha, x_0 + \alpha], \;\; \alpha > 0$, tal que:

(i) $(x, Y(x)) \in D, \;\; \forall x \in I_\alpha$;

(ii) $Y \in C^1(I_\alpha)$ e $Y'(x) = f(x, Y(x)), \;\; \forall x \in I_\alpha$;

(iii) $Y(x_0) = Y_0$.

Esta função é a **solução** do PVI (P).

• Os resultados que iremos obter para o PVI (P) generalizam-se facilmente quer para sistemas de EDO's de 1ª ordem quer para EDO's de ordem superior à 1ª, desde que se utilize a apropriada notação vectorial e matricial. Assim, no caso de um sistema de d EDO's de 1ª ordem, temos:

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(x_0) = Y_0, \end{cases}$$

onde $f : D \subset \mathbb{R}^{1+d} \to \mathbb{R}^d$ e $(x_0, Y_0) \in D$, com

$$\begin{aligned} y &= [y_1 \;\; y_2 \;\; \cdots \;\; y_d]^T \\ f(x, y) &= [f_1(x, y) \;\; f_2(x, y) \;\; \cdots \;\; f_d(x, y)]^T \\ Y_0 &= [Y_{01} \;\; Y_{02} \;\; \cdots \;\; Y_{0d}]^T, \end{aligned}$$

enquanto que no caso de uma EDO de ordem d,

$$\begin{cases} y^{(d)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(d-1)}(x)\right), \\ y(x_0) = Y_0, \;\; y'(x_0) = Y_0', \;\; \ldots, \;\; y^{(d-1)}(x_0) = Y_0^{(d-1)}, \end{cases}$$

temos a formulação equivalente

$$\begin{cases} z'(x) = g(x, z(x)), \\ z(x_0) = Z_0, \end{cases}$$

com

$$\begin{aligned} z &= [y \;\; y' \;\; \cdots \;\; y^{(d-1)}]^T, \\ g(x,z) &= [z_2 \;\; z_3 \;\; \cdots \;\; z_d \;\; f(x, z_1, z_2, \ldots, z_d)]^T, \\ Z_0 &= [Y_0 \;\; Y_0' \;\; \cdots \;\; Y_0^{(d-1)}]^T. \end{aligned}$$

Exemplo. Escrever na forma de um PVI para um sistema de EDO's de 1ª ordem o seguinte PVI para uma EDO de 2ª ordem:

$$\begin{cases} y''(x) + a(x)y'(x) + b(x)y(x) = r(x), \\ y(x_0) = Y_0, \qquad y'(x_0) = Y_0', \end{cases}$$

onde $a, b, r : I \subset \mathbb{R} \to \mathbb{R}$ são funções contínuas num intervalo I e $x_0 \in I$.

$$z(x) = \begin{bmatrix} z_1(x) \\ z_2(x) \end{bmatrix} = \begin{bmatrix} y(x) \\ y'(x) \end{bmatrix}$$

$$z'(x) = \begin{bmatrix} y'(x) \\ y''(x) \end{bmatrix} = \begin{bmatrix} z_2(x) \\ r(x) - b(x)z_1(x) - a(x)z_2(x) \end{bmatrix} = g(x, z(x))$$

$$z(x_0) = \begin{bmatrix} y(x_0) \\ y'(x_0) \end{bmatrix} = \begin{bmatrix} Y_0 \\ Y_0' \end{bmatrix} = Z_0$$

• Antes de iniciar a análise numérica do PVI (P) apresentamos alguns resultados teóricos sobre este problema que ajudam a compreender melhor os métodos numéricos que discutiremos a seguir. São tratados com mais pormenor na disciplina de Análise Complexa e Equações Diferenciais.

Proposição (Teorema de Picard-Lindelöf). Seja $f : D \subset \mathbb{R}^{1+1} \to \mathbb{R}$ uma função contínua numa região D e que satisfaz a uma condição de Lipschitz em relação à 2ª variável com constante L em D, isto é,

$$|f(x,u) - f(x,v)| \leq L|u - v|, \qquad \forall (x,u), (x,v) \in D.$$

Então para qualquer $(x_0, Y_0) \in D$ o PVI (P) tem uma solução única definida num intervalo $[x_0 - \alpha, x_0 + \alpha]$ para algum $\alpha > 0$ suficientemente pequeno.

Notas.

(a) Sendo

$$D = \{(x,y) \in \mathbb{R}^{1+1} : \; |x - x_0| \leq a, \;\; |y - Y_0| \leq b\},$$

com $a, b > 0$, então

$$\alpha = \min\left\{a, \frac{b}{M}\right\}, \qquad M = \max_{(x,y)\in D} |f(x,y)|.$$

(b) Sendo
$$D = \{(x, y) \in \mathbb{R}^{1+1} : \ |x - x_0| \leq a, \ \ |y| < \infty\},$$
então $\alpha = a$.

(c) Sendo $D = \mathbb{R}^{1+1}$, f contínua em D e satisfazendo a uma condição de Lipschitz em qualquer conjunto
$$D_a = \{(x, y) \in \mathbb{R}^{1+1} : \ |x| \leq a, \ \ |y| < \infty\},$$
então $\alpha = \infty$.

Notas.

(a) Se $D \subset \mathbb{R}^{1+d}$ o teorema continua válido, sendo agora a condição de Lipschitz
$$\|f(x, u) - f(x, v)\| \leq L\|u - v\|, \qquad \forall (x, u), (x, v) \in D,$$
onde $\|\cdot\|$ designa uma qualquer norma vectorial em $\mathbb{R}^d$.

(b) Se $f \in C^1(D), \quad D \subset \mathbb{R}^{1+1}$, então f satisfaz a uma condição de Lipschitz com constante
$$L = \sup_{(x,y)\in D} \left| \frac{\partial f(x, y)}{\partial y} \right|.$$

(c) Se $f \in C^1(D), \quad D \subset \mathbb{R}^{1+d}$, então f satisfaz a uma condição de Lipschitz com constante
$$L = \sup_{(x,y)\in D} \|J_f(x, y)\|,$$
onde $\|\cdot\|$ é a norma matricial induzida pela norma vectorial utilizada em $\mathbb{R}^d$.

Exemplo. O PVI
$$\begin{cases} y'(x) = [y(x)]^2, \\ y(x_0) = Y_0, \end{cases}$$
onde $x_0 \in \mathbb{R}, Y_0 \in \mathbb{R} \setminus \{0\}$ tem a solução única
$$Y(x) = \frac{1}{Y_0^{-1} - (x - x_0)},$$
definida para $x - x_0 < Y_0^{-1}$, se $Y_0 > 0$, e para $x - x_0 > Y_0^{-1}$, se $Y_0 < 0$.

Exemplo. O PVI
$$\begin{cases} y'(x) = 2\sqrt{y(x)}, \\ y(0) = 0. \end{cases}$$
tem uma infinidade de soluções dadas por
$$Y_c(x) = \begin{cases} 0, & x \leq c, \\ (x - c)^2, & x > c, \end{cases}$$

onde $c \geq 0$. O PVI tem também a solução

$$\tilde{Y}(x) = 0, \qquad x \in \mathbb{R}.$$

Note-se que $f(y) = 2\sqrt{y}$ é contínua em $[0, \infty[$ mas não satisfaz a uma condição de Lipschitz em qualquer intervalo que contenha a origem. Consequentemente o teorema de Picard-Lindelöf não é aplicável.

Exemplo. O PVI

$$\begin{cases} y'(x) = \lambda y(x) + g(x), \\ y(x_0) = Y_0, \end{cases}$$

onde $g \in C(\mathbb{R}),\ \lambda, x_0, Y_0 \in \mathbb{R}$, tem a solução única definida em $\mathbb{R}$,

$$Y(x) = Y_0 e^{\lambda(x-x_0)} + \int_{x_0}^{x} e^{\lambda(x-t)} g(t)\, dt.$$

Introdução: métodos numéricos

• Consideremos o PVI

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(x_0) = Y_0, \end{cases} \tag{P}$$

onde $f : D \subset \mathbb{R}^{1+1} \to \mathbb{R}$ é uma função contínua numa região D e que satisfaz a uma condição de Lipschitz em relação à $2^{\underline{a}}$ variável. Para qualquer $(x_0, Y_0) \in D$ o PVI (P) tem uma solução única $Y : I_\alpha \to \mathbb{R}$, para algum $\alpha > 0$.

Pretendemos obter uma aproximação desta solução única do PVI (P) no intervalo $[x_0, b] \subset I_\alpha$.

Consideremos então uma partição do intervalo $[x_0, b]$:

$$x_0 < x_1 < \cdots < x_N = b,$$

com

$$x_n = x_0 + nh, \qquad n = 0, 1, \ldots, N, \qquad h = \frac{b - x_0}{N},$$

onde $h > 0$ é o **passo** da partição.

O objectivo dos métodos numéricos para a resolução do PVI (P) é construir aproximações $y_0, y_1, \ldots, y_N$ para os valores da solução exacta $Y(x_0), Y(x_1), \ldots, Y(x_N)$ nos pontos da partição $x_0, x_1, \ldots, x_N$.

Definição. Um **método de passo simples** ou **único** para a resolução do PVI (P) consiste em construir uma aproximação y_n para a solução exacta $Y(x_n)$ em cada ponto x_n pela fórmula

$$y_{n+1} = y_n + h\, \varphi(x_{n+1}, y_{n+1}, x_n, y_n; h), \qquad n \geq 0,$$

onde y_0 é tal que $(x_0, y_0) \in D$ e $\varphi : D^2 \times]0, \infty[\to \mathbb{R}$ é uma função dada em termos de f, designada por vezes por *função incremento*. Se φ não depender de y_{n+1} o método diz-se **explícito**; caso contrário o método diz-se **implícito**.

Definição. Um **método de passo múltiplo** ou **multipasso**, com $p+1$ passos, $p \in \mathbb{N}_1$, para a resolução do PVI (P) consiste em construir uma aproximação y_n para a solução exacta $Y(x_n)$ em cada ponto x_n pela fórmula

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h\varphi(x_{n+1}, y_{n+1}, x_n, y_n, x_{n-1}, y_{n-1}, \ldots, x_{n-p}, y_{n-p}; h), \qquad n \geq p,$$

onde $y_0, y_1, \ldots, y_p$ são valores dados, ou obtidos por outro método, tais que $(x_0, y_0), \ldots, (x_p, y_p) \in D$ e $\varphi : D^{p+2} \times]0, \infty[\to \mathbb{R}$. Se φ não depender de y_{n+1} o método diz-se **explícito**; caso contrário o método diz-se **implícito**.

Exemplos. Os quatro primeiros são métodos de passo simples. O quarto e o sexto são métodos implícitos

(i) Método de Euler (ordem 1)

$$y_{n+1} = y_n + hf(x_n, y_n), \qquad n \geq 0.$$

(ii) Método de Taylor de ordem 2

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right)(x_n, y_n), \qquad n \geq 0.$$

(iii) Método de Runge-Kutta clássico de ordem 2

$$y_{n+1} = y_n + \frac{h}{4}\left[f(x_n, y_n) + 3f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}\, f(x_n, y_n)\right)\right], \qquad n \geq 0.$$

(iv) Método trapezoidal ou método de Adams-Moulton com um passo (ordem 2)

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_{n+1}, y_{n+1}) + f(x_n, y_n)\right], \qquad n \geq 0.$$

(v) Método de Adams-Bashforth com dois passos (ordem 2)

$$y_{n+1} = y_n + \frac{h}{2}\left[3f(x_n, y_n) - f(x_{n-1}, y_{n-1})\right], \qquad n \geq 1.$$

(vi) Método de Adams-Moulton com dois passos (ordem 3)

$$y_{n+1} = y_n + \frac{h}{12}\left[5f(x_{n+1}, y_{n+1}) + 8f(x_n, y_n) - f(x_{n-1}, y_{n-1})\right], \qquad n \geq 1.$$

Métodos de passo simples: consistência e convergência

Definição. Um **método de passo simples** ou **único**, **explícito**, para a resolução do PVI (P) consiste em construir uma aproximação y_n para a solução exacta $Y(x_n)$ em cada ponto x_n pela fórmula

$$y_{n+1} = y_n + h\,\varphi(x_n, y_n; h), \qquad n \in \mathbb{N},$$

onde y_0 é tal que $(x_0, y_0) \in D$ e $\varphi : D\times]0, \infty[\to \mathbb{R}$ é uma função dada em termos de f.

Exemplo. "Dedução" do método de Euler: (i) a partir da fórmula de Taylor; (ii) a partir da equação integral equivalente usando uma fórmula de quadratura para aproximar o integral.

(i) Fórmula de Taylor

$$Y(x+h) = Y(x) + hY'(x) + \frac{h^2}{2} Y''(x+\theta h), \qquad \theta \in]0,1[.$$

$$Y(x+h) = Y(x) + hf(x, Y(x)) + \frac{h^2}{2} \left(\frac{d}{dx} f(x, Y(x)) \right) (x+\theta h).$$

(ii) Fórmula de quadratura aplicada à equação integral

$$Y(x+h) = Y(x) + \int_x^{x+h} f(t, Y(t))\, dt,$$

$$\int_x^{x+h} g(t)\, dt = hg(x) + \frac{h^2}{2} g'(x+\theta h),$$

$$Y(x+h) = Y(x) + hf(x, Y(x)) + \frac{h^2}{2} \left(\frac{d}{dx} f(x, Y(x)) \right) (x+\theta h).$$

Desprezando os termos de $\mathcal{O}(h^2)$ obtém-se

$$Y(x+h) \approx Y(x) + hf(x, Y(x)),$$

$$y_{n+1} = y_n + hf(x_n, y_n).$$

A quantidade

$$\tau(x, y; h) = \frac{Y(x+h) - Y(x)}{h} - f(x, Y(x)) = \frac{h}{2} Y''(x+\theta h),$$

designada por *erro de discretização local*, desempenha um papel importante no estudo dos métodos numéricos para EDO's.

Nota. Qualquer destas abordagens pode ser generalizada para obter métodos de ordem mais elevada.

Definição. Para cada $(x, y) \in D$ seja $Z(t)$ a solução única do PVI

$$\begin{cases} z'(t) = f(t, z(t)), \\ z(x) = y. \end{cases}$$

Chama-se **erro de discretização local** a

$$\tau(x, y; h) = \frac{Z(x+h) - Z(x)}{h} - \varphi(x, y; h), \qquad Z(x) = y.$$

O método de passo simples diz-se **consistente** (com o PVI) se

$$\lim_{h\to 0} \tau(x, y; h) = 0,$$

uniformemente para todos os $(x, y) \in D$, e diz-se ter **consistência de ordem** $q \in \mathbb{N}_1$ se

$$\tau(x, y; h) = \mathcal{O}(h^q), \qquad h \to 0,$$

$\forall (x, y) \in D$.

Nota. $F(h) = \mathcal{O}(h^q), \;\; h \to 0+, \;\; q > 0$, se existirem $h_0 > 0$ e $C > 0$ tais que

$$|F(h)| \leq Ch^q, \qquad \forall h \in]0, h_0].$$

Nota. O erro de discretização local é a diferença entre a diferença dividida da solução exacta do PVI e a diferença dividida da solução aproximada do PVI, para o mesmo passo h. É pois uma medida da forma como a solução exacta satisfaz à equação do método numérico.

Proposição. Um método de passo simples é consistente se e só se

$$\lim_{h \to 0} \varphi(x, y; h) = f(x, y),$$

uniformemente para todos os $(x, y) \in D$.

Proposição. O método de Euler é consistente. Se $f \in C^1(D)$ então o método de Euler tem consistência de ordem 1.

Definição. Sejam $y_0, y_1, \ldots, y_N$ os valores aproximados obtidos por um método de passo simples para os valores $Y(x_0), Y(x_1), \ldots, Y(x_N)$ da solução do PVI (P). Chama-se **erro de discretização global** a

$$e_n = e_n(h) := Y(x_n) - y_n, \qquad n = 0, 1, \ldots, N,$$

e chama-se **erro de discretização global máximo** a

$$E = E(h) := \max_{0 \leq n \leq N} |e_n(h)|.$$

O método de passo simples diz-se **convergente** se

$$\lim_{h \to 0} E(h) = 0,$$

e diz-se ter **convergência de ordem** q se

$$E(h) = \mathcal{O}(h^q), \qquad h \to 0.$$

Proposição. Considere-se um método de passo simples em que a função φ é contínua e Lipschitziana em relação à segunda variável com constante de Lipschitz M independente de h, isto é,

$$\exists h_0 > 0, \;\; \exists M > 0 : \;\; \forall h \in]0, h_0], \forall (x, y), (x, z) \in D \Rightarrow |\varphi(x, y; h) - \varphi(x, z; h)| \leq M|y - z|.$$

Então o erro de discretização global satisfaz à desigualdade

$$|e_n(h)| \leq e^{M(x_n-x_0)}|e_0(h)| + \frac{\tau(h)}{M}\left[e^{M(x_n-x_0)}-1\right], \qquad 0 \leq n \leq N(h),$$

e o erro de discretização global máximo satisfaz à desigualdade

$$E(h) \leq e^{M(b-x_0)}|e_0(h)| + \frac{\tau(h)}{M}\left[e^{M(b-x_0)}-1\right],$$

onde

$$\tau(h) = \max_{0\leq n\leq N(h)-1} |\tau(x_n, Y(x_n); h)|.$$

Se o método for consistente e $e_0(h) \to 0,\ h \to 0$, então o método é convergente. Se o método tiver consistência de ordem q e $e_0(h) = O(h^q),\ h \to 0$, então o método tem convergência de ordem q.

Dem.: $(\cdots)$

Métodos de passo simples: métodos de Taylor

• Vimos como o método de Euler pode ser "deduzido" a partir da fórmula de Taylor

$$Y(x+h) = Y(x) + hY'(x) + \frac{h^2}{2}Y''(x+\theta h), \qquad \theta \in]0,1[,$$

considerando a aproximação

$$Y(x+h) \approx Y(x) + hY'(x),$$

e usando a EDO para exprimir a derivada $Y'(x)$ em termos de $Y(x)$:

$$Y'(x) = f(x, Y(x)).$$

Considerando a fórmula de Taylor de ordem q

$$Y(x+h) = \sum_{j=0}^{q} \frac{h^j}{j!}Y^{(j)}(x) + \frac{h^{q+1}}{(q+1)!}Y^{(q+1)}(x+\theta h), \qquad \theta \in]0,1[,$$

e procedendo de forma semelhante obtemos os métodos de Taylor de ordem q. As sucessivas derivadas $Y^{(j)}(x)$ são obtidas a partir da EDO:

$$\begin{aligned} Y''(x) &= f_x(x, Y(x)) + f_y(x, Y(x))Y'(x) \\ &= f_x(x, Y(x)) + f_y(x, Y(x))f(x, Y(x)) \\ &= (d_f f)(x, Y(x)) \end{aligned}$$

$$Y^{(j)}(x) = (d_f^{j-1} f)(x, Y(x)), \qquad j \geq 3,$$

onde d_f designa o operador diferencial parcial definido por

$$d_f g = g_x + f g_y = \frac{\partial g}{\partial x} + f \frac{\partial g}{\partial y}.$$

para qualquer função g diferenciável em D.

Definição. Os **métodos de Taylor de ordem** q são métodos de passo simples em que a função de incremento é dada por

$$\varphi(x, y; h) = \sum_{k=0}^{q-1} \frac{h^k}{(k+1)!} \left(d_f^k f\right)(x, y).$$

Proposição. O método de Taylor de ordem $q \in \mathbb{N}_1$ é consistente. Se $f \in C^q(D)$ então o método tem consistência de ordem q.

Dem.: $(\cdots)$

Proposição. O método de Taylor de ordem $q \in \mathbb{N}_1$ é convergente. Se $f \in C^q(D)$ então o método tem convergência de ordem q.

Dem.: $(\cdots)$

Métodos de passo simples: métodos de Runge-Kutta

• Os métodos de Runge-Kutta imitam os métodos de Taylor de ordem q mas requerem apenas o cálculo de valores da função f e não das suas derivadas.

Definição. Os **métodos de Runge-Kutta de** s **etapas, explícitos**, são métodos de passo simples em que a função incremento é

$$\varphi(x, y; h) = \sum_{j=1}^{s} \gamma_j \varphi_j(x, y; h),$$

onde

$$\begin{cases} \varphi_1(x, y; h) = f(x, y) \\ \varphi_j(x, y; h) = f\left(x + \alpha_j h, y + h \sum_{i=1}^{j-1} \beta_{ji} \varphi_i(x, y; h)\right), \qquad 2 \leq j \leq s. \end{cases}$$

Os coeficientes são determinados por forma a tornar a ordem de consistência, e, portanto, a ordem de convergência, a maior possível. A ordem de consistência q é a ordem $(\mathcal{O}(h^q), \ \ h \to 0)$ do erro de discretização local:

$$\tau(x, y; h) = \frac{Z(x+h) - y}{h} - \varphi(x, y; h), \qquad Z(x) = y.$$

Proposição. (Teorema de Butcher)

(1) $s \geq q$; (2) $s > q \geq 5$.

Nota.

s	1	2	3	4	5	6	7	8	$\cdots$
$q_{\max}$	1	2	3	4	4	5	6	6	$\cdots$
$\nu(s)$	1	4	8	13	19	26	34	43	$\cdots$

$$\nu(s) = \frac{s(s+3)}{2} - 1 \qquad \text{(número de coeficientes)}$$

Quadro de Butcher:

$$\begin{array}{c|ccccc} \alpha_1 & & & & & \\ \alpha_2 & \beta_{21} & & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \alpha_q & \beta_{q1} & \beta_{q2} & \cdots & \beta_{q,q-1} & \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_{q-1} & \gamma_q \end{array}$$

Exemplo. Os seguintes métodos de Runge-Kutta encontram-se no Anexo. Apresentam-se aqui os respectivos quadros de Butcher.

◇ Métodos de Runge-Kutta de ordem 2 ($s = 2$)
- Método de Euler modificado ou do ponto médio
- Método de Runge-Kutta clássico
- Método de Heun

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ \frac{2}{3} & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

◇ Métodos de Runge-Kutta de ordem 3 ($s = 3$)
- Método de Runge-Kutta clássico
- Método de Runge-Kutta-Nystrom

– Método de Runge-Kutta-Heun

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

◇ Métodos de Runge-Kutta de ordem 4 ($s = 4, s = 5$)

– Método de Runge-Kutta clássico ($s = 4$)

– Método de Runge-Kutta-Gill ($s = 4$)

– Método de Runge-Kutta-Merson ($s = 5$)

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & \frac{\sqrt{2}-1}{2} & \frac{2-\sqrt{2}}{2} & & & \\ 1 & 0 & -\frac{\sqrt{2}}{2} & \frac{2+\sqrt{2}}{2} & & \\ \hline & \frac{1}{6} & \frac{2-\sqrt{2}}{6} & \frac{2+\sqrt{2}}{6} & \frac{1}{6} \end{array}$$

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{3} & \frac{1}{3} & & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & \\ \frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & \\ 1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & \\ \hline & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \end{array}$$

Proposição. Os métodos de Runge-Kutta de ordem $q = 2, 3, 4$ considerados são consistentes e convergentes. Se $f \in C^q(D)$ então os métodos têm consistência e convergência de ordem q.

Dem.: $(\cdots)$

Exemplo. Considere os métodos de Runge-Kutta de 2 etapas,

$$\varphi(x,y;h) = \gamma_1 f(x,y) + \gamma_2 f\left(x+\alpha h, y+\beta h f(x,y)\right)$$

Determine os parâmetros $\alpha, \beta, \gamma_1, \gamma_2$ por forma a que os métodos tenham a maior ordem de consistência possível.

$$\tau(x,y;h) = \frac{Z(x+h)-y}{h} - \varphi(x,y;h) \qquad (y = Z(x))$$

$$\begin{aligned} Z(x+h) &= Z(x) + hZ'(x) + \frac{h^2}{2}Z''(x) + \frac{h^3}{6}Z'''(x) + \mathcal{O}(h^4) \\ &= Z(x) + hf(x,Z(x)) + \frac{h^2}{2}(d_f f)(x,Z(x)) + \frac{h^3}{6}(d_f^2 f)(x,Z(x)) + \mathcal{O}(h^4) \end{aligned}$$

$$d_f f = f_x + f f_y$$

$$d_f^2 f = f_{xx} + f_x f_y + 2 f f_{xy} + f f_y^2 + f^2 f_{yy}$$

$$\varphi(x,y;h) = \varphi(x,y;0) + h\frac{\partial\varphi}{\partial h}(x,y;0) + \frac{h^2}{2}\frac{\partial^2\varphi}{\partial h^2}(x,y;0) + \mathcal{O}(h^3)$$

$$\begin{aligned} \tau(x,y;h) = \; & f(x,y) - \varphi(x,y;0) + \frac{h}{2}\left[(d_f f)(x,y) - 2\frac{\partial\varphi}{\partial x}(x,y;0)\right] \\ & + \frac{h^2}{6}\left[(d_f^2 f)(x,y) - 3\frac{\partial^2\varphi}{\partial h^2}(x,y;0)\right] + \mathcal{O}(h^3) \end{aligned}$$

$$\varphi(x,y;0) = (\gamma_1+\gamma_2) f(x,y)$$

$$\frac{\partial\varphi}{\partial h}(x,y;0) = \gamma_2[\alpha f_x + \beta f f_y](x,y)$$

$$\frac{\partial^2\varphi}{\partial h^2}(x,y;0) = \gamma_2[\alpha^2 f_{xx} + 2\alpha\beta f f_{xy} + \beta^2 f^2 f_{yy}](x,y)$$

$$\begin{aligned} \tau(x,y;h) = \; & (1-\gamma_1-\gamma_2) f(x,y) + h\left[\left(\frac{1}{2}-\alpha\gamma_2\right) f_x + \left(\frac{1}{2}-\beta\gamma_2\right) f f_y\right](x,y) \\ & + \frac{h^2}{6}\left[\left(1-3\gamma_2\alpha^2\right) f_{xx} + 2\left(1-3\gamma_2\alpha\beta\right) f f_{xy} + \left(1-3\gamma_2\beta^2\right) f^2 f_{yy}\right. \\ & \left. + f_x f_y + f f_y^2\right](x,y) + \mathcal{O}(h^3) \end{aligned}$$

$$\boxed{\gamma_1 = 1-\gamma_2, \qquad \alpha = \beta = \frac{1}{2\gamma_2}}$$

$$\boxed{\varphi(x,y;h) = (1-\gamma_2) f(x,y) + \gamma_2 f\left(x + \frac{h}{2\gamma_2}, y + \frac{h}{2\gamma_2} f(x,y)\right)}$$

$$\begin{aligned} \tau(x,y;h) &= \frac{h^2}{6}\left[\left(\left(1-\frac{3}{4\gamma_2}\right)(f_{xx} + 2 f f_{xy} + f^2 f_{yy}) + f_x f_y + f f_y^2\right](x,y) + \mathcal{O}(h^3)\right. \\ &= c(f,\gamma_2)h^2 + \mathcal{O}(h^3) \end{aligned}$$

$$|c(f,\gamma_2)| \le \frac{1}{6}\left[\left|1-\frac{3}{4\gamma_2}\right| |f_{xx} + 2 f f_{xy} + f^2 f_{yy}| + \left|f_x f_y + f f_y^2\right|\right]$$

Casos particulares:

– Método de Euler modificado ou do ponto médio ($\gamma_2 = 1$):

$$\varphi(x, y; h) = f\left(x + \frac{h}{2}, y + \frac{h}{2} f(x, y)\right)$$

– Método de Runge-Kutta clássico de ordem 2 $\left(\gamma_2 = \frac{3}{4}\right)$:

$$\varphi(x, y; h) = \frac{1}{4}\left[f(x, y) + 3f\left(x + \frac{2h}{3}, y + \frac{2h}{3} f(x, y)\right)\right]$$

– Método de Heun $\left(\gamma_2 = \frac{1}{2}\right)$:

$$\varphi(x, y; h) = \frac{1}{2}\left[f(x, y) + f\left(x + h, y + hf(x, y)\right)\right]$$

Exemplo. Considere o problema de valor inicial

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(x_0) = y_0, \end{cases}$$

onde $f : I \times \mathbb{R} \to \mathbb{R}$ é contínua e Lipschitziana em relação à segunda variável, I é um intervalo de $\mathbb{R}$, $x_0 \in I$, e y_0 é uma constante real.

(a) Obtenha um valor aproximado y_2^E para $Y(x_0 + h)$ usando dois passos de comprimento $h/2$ do método de Euler.

(b) Obtenha um valor aproximado y_1^T para $Y(x_0 + h)$ usando um passo de comprimento h do método de Taylor de $2^{\underline{a}}$ ordem.

(c) Obtenha um valor aproximado y_1^{RK} para $Y(x_0 + h)$ usando um passo de comprimento h de qualquer dos métodos de Runge-Kutta de $2^{\underline{a}}$ ordem ($s = 2$).

(d) Particularize os resultados das alíneas anteriores para $f(x, y) = x^2 - y$. Tendo em atenção que a solução exacta do problema de valor inicial é,

$$Y(x) = x^2 - 2x + 2 + Ke^{x_0 - x}, \quad \text{onde } K = y_0 - x_0^2 + 2x_0 - 2,$$

obtenha expansões em série de Mac-Laurin de potências de h para os erros

$$Y(x_0 + h) - y_2^E, \qquad Y(x_0 + h) - y_1^T, \qquad Y(x_0 + h) - y_1^{RK}.$$

Resolução:

(a) $y_1 = y_0 + \frac{h}{2} f(x_0, y_0)$

$$y_2^E = y_1 + \frac{h}{2} f(x_1, y_1), \qquad x_1 = x_0 + \frac{h}{2}$$

$$y_2^E = y_0 + \frac{h}{2}\left[f(x_0,y_0) + f\left(x_0+\frac{h}{2}, y_0+\frac{h}{2}f(x_0,y_0)\right)\right]$$

(b) $y_1^T = y_0 + hf(x_0,y_0) + \dfrac{h^2}{2}(d_f f)(x_0,y_0)$

$$d_f f = f_x + f f_y$$

(c) $y_1^{RK} = y_0 + h\left[(1-\gamma)f(x_0,y_0) + \gamma f\left(x_0+\frac{h}{2\gamma}, y_0+\frac{h}{2\gamma}f(x_0,y_0)\right)\right]$

(d) $y_2^E = y_0 + h(x_0^2 - y_0) + \dfrac{h^2}{4}(y_0 - x_0^2 + 2x_0) + \dfrac{h^3}{8}$

$$y_1^T = y_0 + h(x_0^2 - y_0) + \frac{h^2}{2}(y_0 - x_0^2 + 2x_0)$$

$$y_1^{RK} = y_0 + h(x_0^2 - y_0) + \frac{h^2}{2}(y_0 - x_0^2 + 2x_0) + \frac{h^3}{2\gamma}$$

$$Y(x_0+h) - y_2^E = \frac{h^2}{4}(K+2) + \mathcal{O}(h^3)$$

$$Y(x_0+h) - y_1^T = -\frac{h^3}{6}K + \mathcal{O}(h^4)$$

$$Y(x_0+h) - y_1^{RK} = -\frac{h^3}{12}\left(\frac{3}{\gamma} + 2K\right) + \mathcal{O}(h^4)$$

Exemplo. Considere o problema de valor inicial

$$\begin{cases} y'(x) = f(x, y(x), z(x)), \\ z'(x) = g(x, y(x), z(x)), \\ y(x_0) = y_0, \qquad z(x_0) = z_0, \end{cases}$$

onde $f, g : I \times \mathbb{R}^2 \to \mathbb{R}$ são contínuas e Lipschitzianas em relação às segunda e terceira variáveis, I é um intervalo de $\mathbb{R}$, $x_0 \in I$, e y_0, z_0 são constantes reais. Repita as três primeiras alíneas do exemplo anterior.

Resolução:

$$W(x) = \begin{bmatrix} y(x) \\ z(x) \end{bmatrix}, \qquad W'(x) = \begin{bmatrix} y'(x) \\ z'(x) \end{bmatrix} = \begin{bmatrix} f(x,y(x),z(x)) \\ g(x,y(x),z(x)) \end{bmatrix} = F(x, W(x))$$

$$W(x_0) = \begin{bmatrix} y(x_0) \\ z(x_0) \end{bmatrix} = \begin{bmatrix} y_0 \\ z_0 \end{bmatrix} = W_0$$

(a) $W_1 = W_0 + \dfrac{h}{2}F(x_0, W_0) = \begin{bmatrix} y_0 + \frac{h}{2}f(x_0,y_0,z_0) \\ z_0 + \frac{h}{2}g(x_0,y_0,z_0) \end{bmatrix} = \begin{bmatrix} y_1 \\ z_1 \end{bmatrix}$

$$W_2^E = W_1 + \frac{h}{2}\,F(x_1, W_1) = \begin{bmatrix} y_1 + \frac{h}{2}\,f(x_1, y_1, z_1) \\ z_1 + \frac{h}{2}\,g(x_1, y_1, z_1) \end{bmatrix}, \qquad x_1 = x_0 + \frac{h}{2}$$

(b) $W_1^T = W_0 + hF(x_0, W_0) + \dfrac{h^2}{2}\,(d_F F)(x_0, W_0)$

$$d_F F = \left(\frac{\partial}{\partial x} + F \cdot \nabla_W\right) F = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y} + g\frac{\partial}{\partial z}\right) F$$

$$W_1^T = \begin{bmatrix} y_0 + hf(x_0, y_0, z_0) + \frac{h^2}{2}\,(f_x + ff_y + gf_z)(x_0, y_0, z_0) \\ z_0 + hg(x_0, y_0, z_0) + \frac{h^2}{2}\,(g_x + fg_y + gg_z)(x_0, y_0, z_0) \end{bmatrix}$$

(c) $W_1^{RK} = W_0 + h\left[(1-\gamma)F(x_0, W_0) + \gamma F\left(x_0 + \dfrac{h}{2\gamma}, W_0 + \dfrac{h}{2\gamma}\,F(x_0, W_0)\right)\right]$

$$\tilde{x}_1 = x_0 + \frac{h}{2\gamma}$$

$$\tilde{W}_1 = W_0 + \frac{h}{2\gamma}\,F(x_0, W_0) = \begin{bmatrix} y_0 + \frac{h}{2\gamma}\,f(x_0, y_0, z_0) \\ z_0 + \frac{h}{2\gamma}\,g(x_0, y_0, z_0) \end{bmatrix} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{z}_1 \end{bmatrix}$$

$$W_1^{RK} = \begin{bmatrix} y_0 + h\left[(1-\gamma)f(x_0, y_0, z_0) + \gamma f(\tilde{x}_1, \tilde{y}_1, \tilde{z}_1)\right] \\ z_0 + h\left[(1-\gamma)g(x_0, y_0, z_0) + \gamma g(\tilde{x}_1, \tilde{y}_1, \tilde{z}_1)\right] \end{bmatrix}$$

Exemplo. Considere o problema de valor inicial

$$\begin{cases} y''(x) = g(x, y(x), y'(x)), \\ y(x_0) = y_0, \qquad y'(x_0) = z_0, \end{cases}$$

onde $g : I \times \mathbb{R}^2 \to \mathbb{R}$ é contínua e Lipschitziana em relação às segunda e terceira variáveis, I é um intervalo de $\mathbb{R}$, $x_0 \in I$, e y_0, z_0 são constantes reais. Repita as três primeiras alíneas dos exemplos anteriores.

Resolução:

$$W(x) = \begin{bmatrix} y(x) \\ z(x) \end{bmatrix}, \qquad W'(x) = \begin{bmatrix} y'(x) \\ y''(x) \end{bmatrix} = \begin{bmatrix} z(x) \\ g(x, y(x), z(x)) \end{bmatrix} = F(x, W(x))$$

$$W(x_0) = \begin{bmatrix} y(x_0) \\ z(x_0) \end{bmatrix} = \begin{bmatrix} y_0 \\ z_0 \end{bmatrix} = W_0$$

Caso anterior com $f(x, y, z) = z$.

(a) $W_1 = \begin{bmatrix} y_0 + \frac{h}{2}\,z_0 \\ z_0 + \frac{h}{2}\,g(x_0, y_0, z_0) \end{bmatrix} = \begin{bmatrix} y_1 \\ z_1 \end{bmatrix}$

$$W_2^E = \begin{bmatrix} y_1 + \frac{h}{2}\, z_1 \\ z_1 + \frac{h}{2}\, g(x_1, y_1, z_1), \end{bmatrix}, \qquad x_1 = x_0 + \frac{h}{2}$$

(b) $$W_1^T = \begin{bmatrix} y_0 + hz_0 + \frac{h^2}{2}\, g(x_0, y_0, z_0) \\ z_0 + hg(x_0, y_0, z_0) + \frac{h^2}{2}\, (g_x + zg_y + gg_z)(x_0, y_0, z_0) \end{bmatrix}$$

(c) $$\tilde{x}_1 = x_0 + \frac{h}{2\gamma}, \qquad \tilde{W}_1 = \begin{bmatrix} y_0 + \frac{h}{2\gamma}\, z_0 \\ z_0 + \frac{h}{2\gamma}\, g(x_0, y_0, z_0) \end{bmatrix} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{z}_1 \end{bmatrix}$$

$$W_1^{RK} = \begin{bmatrix} y_0 + h\,[(1-\gamma)z_0 + \gamma\tilde{z}_1] \\ z_0 + h\,[(1-\gamma)g(x_0, y_0, z_0) + \gamma g(\tilde{x}_1, \tilde{y}_1, \tilde{z}_1)] \end{bmatrix}$$

Métodos multipasso lineares: introdução

Definição. Um **método multipasso linear** (MPL) com $p+1$ passos ($p \geq 0$) para a resolução do PVI (P) consiste em construir uma aproximação y_n para a solução exacta $Y(x_n)$ em cada ponto $x_n = x_0 + nh,\ n = 0, 1, \ldots, N$, pela fórmula

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k f(x_{n-k}, y_{n-k}), \qquad n \geq p,$$

onde $y_0, y_1, \ldots, y_p$ são valores dados (ou obtidos por outro método) e $a_0, a_1, \ldots, a_p$, $b_{-1}, b_0, b_1, \ldots, b_p$ são constantes reais tais que $|a_p| + |b_p| \neq 0$. Se $b_{-1} = 0$ o método diz-se **explícito**; se $b_{-1} \neq 0$ o método diz-se **implícito**.

• Uma classe importante de métodos MPL baseia-se na integração numérica. A ideia geral é a seguinte. Reescreve-se a EDO como uma equação integral

$$Y(x_{n+1}) = Y(x_{n-r}) + \int_{x_{n-r}}^{x_{n+1}} f(t, Y(t))\, dt,$$

para algum $r \geq 0$ e $n \geq r$; constrói-se um polinómio interpolador de grau $\nu \leq p+1$ para f e substitui-se f por este polinómio no integral; desprezando o erro de quadratura chega-se ao método pretendido. Entre os métodos obtidos por esta forma temos:

- ◇ Métodos de Adams: $r = 0,\ \ \nu = p$ ou $\nu = p+1$
  - – Métodos de Adams explícitos ou de Adams-Bashforth: $\nu = p,\ \ p \geq 0$
  - – Métodos de Adams implícitos ou de Adams-Moulton: $\nu = p+1,\ \ p \geq -1$
- ◇ Métodos de Nyström: $r = 1,\ \ \nu = p,\ \ p \geq 1$
- ◇ Métodos de Milne-Thomson: $r = 1,\ \ \nu = p+1,\ \ p \geq 1$

Iremos considerar em detalhe os métodos de Adams que são os métodos MPL mais usados. Em particular são usados para produzir algoritmos preditor-corrector em que o erro é controlado por variação do passo h e da ordem do método, simultâneamente. O nosso

estudo dos métodos de Adams será no entanto feito como caso particular dos métodos MPL acima definidos e não a partir da integração numérica.

Métodos MPL: consistência

• A consistência é uma propriedade que procura avaliar em que medida a solução exacta do PVI satisfaz à equação do método numérico para a sua resolução.

Definição. Para cada $(x, y) \in D$ seja $Y(t)$ a solução única do PVI

$$\begin{cases} z'(t) = f(t, z(t)), \\ z(x) = y. \end{cases}$$

Chama-se **erro de discretização local** a

$$\tau(x, y; h) = \frac{1}{h}\left[Z(x+h) - \sum_{k=0}^{p} a_k Z(x-kh)\right] - \sum_{k=-1}^{p} b_k f(x-kh, Z(x-kh)).$$

O método MPL diz-se **consistente** (com o PVI) se

$$\lim_{h\to 0} \tau(x, y; h) = 0,$$

uniformemente para todos os $(x, y) \in D$, e diz-se ter **consistência de ordem** $q \in \mathbb{N}_1$ se

$$\tau(x, y; h) = \mathcal{O}(h^q), \qquad h \to 0,$$

$\forall (x, y) \in D$.

Proposição. Sejam $C_0, C_1, \ldots$ as quantidades definidas em termos dos parâmetros de um método MPL por:

$$C_0 = 1 - \sum_{k=0}^{p} a_k, \qquad C_1 = 1 + \sum_{k=0}^{p} k a_k - \sum_{k=-1}^{p} b_k,$$
$$C_j = 1 - \sum_{k=0}^{p} (-k)^j a_k - j \sum_{k=-1}^{p} (-k)^{j-1} b_k, \qquad j = 2, 3, \ldots$$

Então:

(1) Um método MPL é consistente com o PVI (P) para qualquer $f \in C^1(D)$ se e só se

$$C_0 = C_1 = 0.$$

(2) Um método MPL tem consistência de ordem $q \geq 1$ para qualquer $f \in C^{q+1}(D)$ se e só se

$$C_0 = C_1 = C_2 = \cdots = C_q = 0, \qquad C_{q+1} \neq 0.$$

O erro de discretização local tem a forma

$$\tau(x, y; h) = h^q \frac{C_{q+1}}{(q+1)!} (d_f^q f)(x, y) + \mathcal{O}(h^{q+1})$$

Dem.: Obtém-se o desenvolvimento:

$$\tau(x,y;h) = C_0 Z(x) h^{-1} + \sum_{j=1}^{q+1} \frac{C_j}{j!} Z^{(j)}(x)\, h^{j-1} + \mathcal{O}(h^{q+1})$$

Métodos MPL: métodos de Adams

Definição. Os **métodos de Adams** são métodos MPL da forma

$$y_{n+1} = y_n + h \sum_{k=p_0}^{p} b_k f(x_{n-k}, y_{n-k}), \qquad n \geq p,$$

onde $p_0 = 0$ (métodos explícitos) ou $p_0 = -1$ (métodos implíctos) e os $p + 1 - p_0 = q$ parâmetros b_k são unicamente determinados pelo sistema de q equações

$$C_1 = C_2 = \cdots = C_q = 0,$$

onde

$$C_1 = 1 - \sum_{k=p_0}^{p} b_k, \qquad C_j = 1 - j \sum_{k=p_0}^{p} (-k)^{j-1} b_k, \quad j = 2, \ldots, q.$$

Proposição. O método de Adams com $p+1$ passos tem consistência de ordem $q = p+1-p_0$, isto é,

(1) $q = p + 1$, se é explícito;

(2) $q = p + 2$, se é implícito.

Exemplos. Apresentam-se no Anexo os primeiros métodos de Adams. Aqui apresentam-se quadros resumos dos coeficientes e da constante que ocorre no termo de ordem mais baixa do erro de discretização local.

Métodos de Adams-Bashforth $(p_0 = 0)$							
p	q	b_0	b_1	b_2	b_3	b_4	$C_{q+1}/(q+1)!$
1	2	$\frac{3}{2}$	$-\frac{1}{2}$				$\frac{5}{12}$
2	3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$			$\frac{3}{8}$
3	4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$		$\frac{251}{720}$
4	5	$\frac{1901}{720}$	$-\frac{2774}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$	$\frac{95}{288}$

Métodos de Adams-Moulton $(p_0 = -1)$							
p	q	b_{-1}	b_0	b_1	b_2	b_3	$C_{q+1}/(q+1)!$
0	2	$\frac{1}{2}$	$\frac{1}{2}$				$-\frac{1}{12}$
1	3	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$			$-\frac{1}{24}$
2	4	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$		$-\frac{19}{720}$
3	5	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$	$-\frac{3}{160}$

Métodos MPL: convergência

• Consideremos o PVI (P)

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(x_0) = Y_0, \end{cases} \tag{P}$$

onde f é contínua e Lipschitziana em relação a y.

Consideremos o método MPL (M) para resolução numérica aproximada de (P):

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k f(x_{n-k}, y_{n-k}), \qquad p \le n \le N(h) - 1, \tag{M}$$

onde $h_0 \in]0, h_0]$ e h_0 é suficientemente pequeno por forma a que a Eq. (M) tenha a solução $\{y_n(h)\}_{0 \le n \le N(h)}$, $\forall h \in]0, h_0]$.

Definição. Sejam $\{y_n(h)\}_{0 \le n \le N(h)}$ a solução obtida pelo método MPL (M) e $Y : [x_0, b] \to \mathbb{R}$ a solução exacta do PVI (P). Define-se o **erro de discretização global** por

$$e_n = e_n(h) := Y(x_n) - y_n(h), \qquad 0 \le n \le N(h),$$

e **erro de discretização global máximo** por

$$E = E(h) := \max_{0 \le n \le N(h)} |e_n(h)|.$$

Diz-se que a solução aproximada é **convergente** para a solução exacta de

$$\lim_{h \to 0} E(h) = 0,$$

e diz-se que tem **convergência de ordem** q se

$$E(h) = \mathcal{O}(h^q), \qquad h \to 0.$$

Diz-se que o método numérico (M) é convergente, ou tem convergência de ordem q, se a convergência (ou a convergência de ordem q) se verificar para todos os PVI (P).

Definição. Os **valores iniciais** $\{y_n(h)\}_{0 \le n \le p}$ dizem-se **consistentes** se,

$$\lim_{h \to 0} E_p(h) = 0,$$

onde

$$E_p(h) := \max_{0 \le n \le p} |e_n(h)|,$$

e diz-se terem **consistência de ordem** q se

$$E_p(h) = \mathcal{O}(h^q), \qquad h \to 0.$$

Nota. A consistência dos valores iniciais é condição necessária para a convergência do método.

• A convergência do método MPL está relacionada com as raízes do polinómio

$$\rho(r) := r^{p+1} - \sum_{k=0}^{p} a_k r^{p-k}.$$

Definição. Diz-se que o método MPL satisfaz à **condição da raiz** se as raízes $r_0, r_1, \ldots, r_p$ do polinómio $\rho(r)$, repetidas de acordo com as suas multiplicidades, satisfazem a

$$\text{(i)} \ \ |r_j| \le 1, \ \ j = 0, 1, \ldots, p; \qquad \text{(ii)} \ \ |r_j| = 1 \ \ \Rightarrow \ \ \rho'(r_j) \neq 0$$

Nota. A condição (i) impõe que todas as raízes estão contidas no círculo unitário $\{z \in \mathbb{C} : |z| \le 1\}$. A condição (ii) impõe que todas as raízes na fronteira do círculo são raízes simples de $\rho(r)$.

Nota. Fazendo $h = 0$ em (M) obtemos

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k},$$

que é uma equação às diferenças de ordem $p + 1$, linear e de coeficientes constantes. O seu polinómio característico é precisamente ρ. A sua solução geral é dada por

$$y_n = \sum_{j=0}^{\hat{p}} \left(\sum_{l=0}^{\mu_{j-1}} c_{jl} n^l \right) \hat{r}_j^n,$$

onde $\hat{r}_0, \hat{r}_1, \ldots, \hat{r}_{\hat{p}}$, são as raízes distintas do polinómio ρ e $\mu_0, \mu_1, \ldots, \mu_{\hat{p}}$ as suas multiplicidades. É claro que $\hat{p} \le p$ e $\mu_0 + \mu_1 + \cdots + \mu_{\hat{p}} = p + 1$.

A condição da raiz corresponde a exigir que todas as soluções da equação às diferenças são limitadas.

Proposição. No caso de um método MPL consistente as seguintes afirmações são equivalentes:

(1) o método é convergente;

(2) o método satisfaz à condição da raiz.

Proposição. Todo o método de passo simples linear consistente é convergente.

Dem.: $(\cdots)$

Proposição. Os métodos de Adams-Bashforth e Adams-Moulton são convergentes.

Dem.: $(\cdots)$

Exemplo. Para cada um dos três problemas de valor inicial considerados nos exemplos no fim do parágrafo sobre os métodos de Runge-Kutta obtenha um valor aproximado y_1^{AM} para $Y(x_0+h)$ usando um passo de comprimento h do método de Adams-Moulton de $2^{\underline{a}}$ ordem.

Resolução:

(i) $y_1^{AM} = y_0 + \frac{h}{2}\left[f(x_0,y_0) + f\left(x_0+h, y_1^{AM}\right)\right]$

(ii) $W_1^{AM} = W_0 + \frac{h}{2}\left[F(x_0,W_0) + F\left(x_0+h, W_1^{AM}\right)\right]$

$$\begin{bmatrix} y_1^{AM} \\ z_1^{AM} \end{bmatrix} = \begin{bmatrix} y_0 + \frac{h}{2}\left[f(x_0,y_0,z_0) + f\left(x_1, y_1^{AM}, z_1^{AM}\right)\right] \\ z_0 + \frac{h}{2}\left[g(x_0,y_0,z_0) + g\left(x_1, y_1^{AM}, z_1^{AM}\right)\right] \end{bmatrix}$$

(iii) $$\begin{bmatrix} y_1^{AM} \\ z_1^{AM} \end{bmatrix} = \begin{bmatrix} y_0 + \frac{h}{2}\left[z_0 + z_1^{AM}\right] \\ z_0 + \frac{h}{2}\left[g(x_0,y_0,z_0) + g\left(x_1, y_1^{AM}, z_1^{AM}\right)\right] \end{bmatrix}$$

Exemplo. Determinar todos os métodos MPL com 2 passos e ordem de consistência 2.

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + h[b_{-1} f_{n+1} + b_0 f_n + b_1 f_{n-1}]$$

Resolução:

(i) Condições para que o método tenha ordem de consistência ≥ 2:

$$\begin{cases} C_0 = 1 - a_0 - a_1 = 0 \\ C_1 = 1 + a_1 - b_{-1} - b_0 - b_1 = 0 \\ C_2 = 1 - a_1 - 2(b_{-1} - b_1) = 0 \end{cases} \Rightarrow \begin{cases} a_1 = 1 - a_0 \\ b_0 = 2 - \frac{a_0}{2} - 2b_{-1} \\ b_1 = -\frac{a_0}{2} + b_{-1} \end{cases}$$

$$\Rightarrow \begin{cases} C_3 = 1 + a_1 - 3(b_{-1} + b_1) = 2 + \frac{a_0}{2} - 6b_{-1} \\ C_4 = 1 - a_1 - 4(b_{-1} - b_1) = -a_0 \\ C_5 = 1 + a_1 - 5(b_{-1} + b_1) = 2 + \frac{3a_0}{2} - 10b_{-1} \end{cases}$$

(ii) Condição da raiz:

$\rho(r) = r^2 - a_0 r - a_1 = (r-1)(r+1-a_0)$

$\boxed{0 \leq a_0 < 2}$

(iii) Casos particulares:

$\diamond$ $a_0 = 0, \quad b_{-1} = 0, \quad a_1 = 1, \quad b_0 = 2, \quad b_1 = 0, \quad C_3 = 2$

$$y_{n+1} = y_{n-1} + 2hf_n$$

Método do ponto médio

$\diamond$ $a_0 = 1, \quad b_{-1} = 0, \quad a_1 = 0, \quad b_0 = \frac{3}{2}, \quad b_1 = -\frac{1}{2}, \quad C_3 = \frac{5}{2}$

$$y_{n+1} = y_n + \frac{h}{2}\left[3f_n - f_{n-1}\right]$$

Método de Adams-Bashforth de ordem 2

$\diamond$ $a_0 = \frac{4}{3}, \quad b_{-1} = \frac{2}{3}, \quad a_1 = -\frac{1}{3}, \quad b_0 = 0, \quad b_1 = 0, \quad C_3 = -\frac{4}{3}$

$$y_{n+1} = \frac{4}{3}\,y_n - \frac{1}{3}\,y_{n-1} + \frac{2h}{3}\,f_{n+1}$$

Método BDF de ordem 2 (Backward Difference Formula)

(iv) Condições para que o método tenha ordem de consistência ≥ 3:

$$C_3 = 0 \quad \Leftrightarrow \quad b_{-1} = \frac{1}{12}\,(a_0 + 4)$$

$$\Rightarrow \quad \begin{cases} a_1 = 1 - a_0, & b_0 = \frac{2}{3}\,(2 - a_0), \qquad b_1 = \frac{1}{12}\,(4 - 5a_0), \\ C_4 = -a_0, & C_5 = \frac{2}{3}\,(a_0 - 2) \end{cases}$$

(v) Casos particulares:

$\diamond$ $a_0 = 1, \quad b_{-1} = \frac{5}{12}, \quad a_1 = 0, \quad b_0 = \frac{2}{3}, \quad b_1 = -\frac{1}{12}, \quad C_4 = -1$

$$y_{n+1} = y_n + \frac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right]$$

Método de Adams-Moulton de ordem 3

(vi) Condições para que o método tenha ordem de consistência ≥ 4:

$$C_4 = 0 \quad \Leftrightarrow \quad a_0 = 0$$

$$\Rightarrow \quad b_{-1} = \frac{1}{3}, \quad a_1 = 1, \quad b_0 = \frac{4}{3}, \quad b_1 = \frac{1}{3}, \quad C_5 = -\frac{4}{3}$$

$$y_{n+1} = y_{n-1} + \frac{h}{3}\left[f_{n+1} + 4f_n + f_{n-1}\right]$$

Método de Milne de ordem 4

Métodos MPL: métodos preditor-corrector

• Consideremos a expressão geral dos métodos de Adams

$$y_{n+1} = y_n + h \sum_{k=p_0}^{p} b_k f(x_{n-k}, y_{n-k}), \qquad n \geq p.$$

O valor inicial é um dado do PVI (P), $y_0 = Y_0$. Os valores iniciais $y_1, y_2, \ldots, y_p$ têm que ser obtidos por outras vias, por exemplo, por um método de passo simples (da mesma ordem).

• As fórmulas de Adams-Moulton, que definem implicitamente y_{n+1}, podem ser resolvidas pelo método iterativo do ponto fixo:

$$y_{n+1}^{(j+1)} = y_n + h \sum_{k=0}^{p} b_k f(x_{n-k}, y_{n-k}) + h b_{-1} f(x_{n+1}, y_{n+1}^{(j)}), \quad n \geq p, \;\; j \geq 0.$$

A convergência do método do ponto fixo fica assegurada se h for suficientemente pequeno, de acordo com o seguinte resultado:

Proposição. O método iterativo do ponto fixo aplicado à equação do método de Adams-Moulton converge para a solução y_{n+1} desde que seja satisfeita a desigualdade

$$h|b_{-1}|L < 1,$$

onde h é o passo de integração e L designa a constante de Lipschitz de f.

Dem.: $(\cdots)$

• Para obter a iterada inicial $y_{n+1}^{(0)}$ pode utilizar-se um método de Adams-Bashforth.

Definição. Os **métodos preditor-corrector** são constituídos por uma fórmula de Adams-Moulton (o **corrector**) e uma fórmula de Adams-Bashforth (o **preditor**):

$$\begin{cases} y_{n+1}^{(j+1)} = y_n + h \displaystyle\sum_{k=0}^{p} b_k f(x_{n-k}, y_{n-k}) + h b_{-1} f(x_{n+1}, y_{n+1}^{(j)}), & n \geq p, \;\; j \geq 0, \\ y_{n+1}^{(0)} = y_n + h \displaystyle\sum_{k=0}^{\tilde{p}} \tilde{b}_k f(x_{n-k}, y_{n-k}), & n \geq \tilde{p}. \end{cases}$$

Exemplos.

(i) (AB3)+(AM3)

$$\begin{cases} y_{n+1}^{(0)} = y_n + \dfrac{h}{12}\left[23 f_n - 16 f_{n-1} + 5 f_{n-2}\right], \\ y_{n+1}^{(j+1)} = y_n + \dfrac{h}{12}\left[5 f(x_{n+1}, y_{n+1}^{(j)}) + 8 f_n - f_{n-1}\right], & j \geq 0, \;\; n \geq 1. \end{cases}$$

(ii) (AB2)+(AM3)

$$\begin{cases} y_{n+1}^{(0)} = y_n + \dfrac{h}{2}\,[3f_n - f_{n-1}], \\ y_{n+1}^{(j+1)} = y_n + \dfrac{h}{12}\,\left[5f(x_{n+1}, y_{n+1}^{(j)}) + 8f_n - f_{n-1}\right], \qquad j \geq 0, \;\; n \geq 1. \end{cases}$$

Proposição. Consideremos um preditor de ordem $\tilde{q}$ e um corrector de ordem q.

(1) Se $\tilde{q} \geq q$ então o preditor-corrector tem a mesma ordem e o mesmo EDL principal que o corrector.

(2) Se $\tilde{q} = q - 1$ então o preditor-corrector tem a mesma ordem que o corrector mas o EDL principal é diferente.

(3) Se $\tilde{q} \leq q - 2$ então o preditor-corrector tem ordem mais baixa que o corrector.

Nota. Vimos que o erro de discretização local (EDL) para um método de Adams de ordem q é dado por

$$\tau(x, y; h) = h^q \, \frac{C_{q+1}}{(q+1)!} \, (d_f^q f)(x, y) + \mathcal{O}(h^{q+1})$$

Chama-se EDL principal à primeira parcela do EDL, proporcional a h^q. A mesma designação aplica-se no caso do método preditor-corrector.

Dem.: $(\cdots)$

Anexo

- Métodos de Runge-Kutta de ordem 2:

$$\tau(x,y;h) = \frac{h^2}{6}\left[d_f^2 f(x,y) - 3\,\frac{\partial^2\varphi}{\partial h^2}(x,y;0)\right] + \mathcal{O}(h^3)$$

  - Método de Euler modificado ou do ponto médio

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\,f(x_n,y_n)\right)$$

  - Método de Runge-Kutta clássico de ordem 2

$$y_{n+1} = y_n + \frac{h}{4}\left[f(x_n,y_n) + 3f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}\,f(x_n,y_n)\right)\right]$$

  - Método de Heun

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n,y_n) + f\left(x_n + h, y_n + hf(x_n,y_n)\right)\right]$$

- Métodos de Runge-Kutta de ordem 3:

$$\tau(x,y;h) = \frac{h^3}{24}\left[d_f^3 f(x,y) - 4\,\frac{\partial^3\varphi}{\partial h^3}(x,y;0)\right] + \mathcal{O}(h^4)$$

  - Método de Runge-Kutta clássico de ordem 3

$$y_{n+1} = y_n + \frac{h}{6}\left[\varphi_1 + 4\varphi_2 + \varphi_3\right]$$
$$\varphi_1 = f(x_n,y_n), \qquad \varphi_2 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + h, y_n - h\varphi_1 + 2h\varphi_2\right)$$

  - Método de Runge-Kutta-Nystrom de ordem 3

$$y_{n+1} = y_n + \frac{h}{8}\left[2\varphi_1 + 3\varphi_2 + 3\varphi_3\right]$$
$$\varphi_1 = f(x_n,y_n), \qquad \varphi_2 = f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}\,\varphi_2\right)$$

  - Método de Runge-Kutta-Heun de ordem 3

$$y_{n+1} = y_n + \frac{h}{4}\left[\varphi_1 + 3\varphi_3\right]$$
$$\varphi_1 = f(x_n,y_n), \qquad \varphi_2 = f\left(x_n + \frac{h}{3}, y_n + \frac{h}{3}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}\,\varphi_2\right)$$

- Métodos de Runge-Kutta de ordem 4:

$$\tau(x,y;h) = \frac{h^4}{120}\left[d_f^4 f(x,y) - 5\,\frac{\partial^4\varphi}{\partial h^4}(x,y;0)\right] + \mathcal{O}(h^5)$$

  – Método de Runge-Kutta clássico de ordem 4:

$$y_{n+1} = y_n + \frac{h}{6}\left[\varphi_1 + 2\varphi_2 + 2\varphi_3 + \varphi_4\right]$$

$$\varphi_1 = f(x_n, y_n), \qquad \varphi_2 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\,\varphi_2\right), \qquad \varphi_4 = f(x_n + h, y_n + h\varphi_3)$$

  – Método de Runge-Kutta-Gill de ordem 4:

$$y_{n+1} = y_n + \frac{h}{6}\left[\varphi_1 + (2-\sqrt{2})\varphi_2 + (2+\sqrt{2})\varphi_3 + \varphi_4\right]$$

$$\varphi_1 = f(x_n, y_n), \qquad \varphi_2 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + \frac{h}{2}, y_n + \frac{\sqrt{2}-1}{2}\,h\varphi_1 + \frac{2-\sqrt{2}}{2}\,h\varphi_2\right)$$
$$\varphi_4 = f\left(x_n + h, y_n - \frac{\sqrt{2}}{2}\,h\varphi_2 + \frac{2+\sqrt{2}}{2}\,h\varphi_3\right)$$

  – Método de Runge-Kutta-Merson de ordem 4:

$$y_{n+1} = y_n + \frac{h}{6}\left[\varphi_1 + 4\varphi_4 + \varphi_5\right]$$

$$\varphi_1 = f(x_n, y_n), \qquad \varphi_2 = f\left(x_n + \frac{h}{3}, y_n + \frac{h}{3}\,\varphi_1\right)$$
$$\varphi_3 = f\left(x_n + \frac{h}{3}, y_n + \frac{h}{6}\,\varphi_1 + \frac{h}{6}\,\varphi_2\right)$$
$$\varphi_4 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{8}\,\varphi_1 + \frac{3h}{8}\,\varphi_3\right)$$
$$\varphi_5 = f\left(x_n + h, y_n + \frac{h}{2}\,\varphi_1 - \frac{3h}{2}\,\varphi_3 + 2h\varphi_4\right)$$

- Métodos de Adams-Bashforth ($f_m := f(x_m, y_m)$):

$$\text{(AB2)}\ \ p=1,\ \ q=2\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{2}\left[3f_n - f_{n-1}\right] \\ \tau(x,y;h) = \dfrac{5h^2}{12}\,(d_f^2 f)(x,y) + \mathcal{O}(h^3) \end{cases}$$

$$\text{(AB3)}\ \ p=2,\ \ q=3\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{12}\left[23f_n - 16f_{n-1} + 5f_{n-2}\right] \\ \tau(x,y;h) = \dfrac{3h^3}{8}\,(d_f^3 f)(x,y) + \mathcal{O}(h^4) \end{cases}$$

$$\text{(AB4)}\ \ p=3,\ \ q=4\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{24}\left[55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right], \\ \tau(x,y;h) = \dfrac{251h^4}{720}\,(d_f^4 f)(x,y) + \mathcal{O}(h^5) \end{cases}$$

$$\text{(AB5)}\ \ p=4,\ \ q=5\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{720}\ \left[1901f_n - 2774f_{n-1} + 2616f_{n-2}\right. \\ \qquad\qquad\qquad\qquad \left. -1274f_{n-3} + 251f_{n-4}\right], \\ \tau(x,y;h) = \dfrac{95h^5}{288}\,(d_f^5 f)(x,y) + \mathcal{O}(h^6) \end{cases}$$

- Métodos de Adams-Moulton ($f_m := f(x_m, y_m)$):

$$\text{(AM2)}\ p=0,\ \ q=2\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{2}\left[f_{n+1} + f_n\right] \\ \tau(x,y;h) = -\dfrac{h^2}{12}\,(d_f^2 f)(x,y) + \mathcal{O}(h^3) \end{cases}$$

$$\text{(AM3)}\ \ p=1,\ \ q=3\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right] \\ \tau(x,y;h) = -\dfrac{h^3}{24}\,(d_f^3 f)(x,y) + \mathcal{O}(h^4) \end{cases}$$

$$\text{(AM4)}\ \ p=2,\ \ q=4\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{24}\left[9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right] \\ \tau(x,y;h) = -\dfrac{19h^4}{720}\,(d_f^4 f)(x,y) + \mathcal{O}(h^5) \end{cases}$$

$$\text{(AM5)}\ \ p=3,\ \ q=5\text{:}\ \begin{cases} y_{n+1} = y_n + \dfrac{h}{720}\ \left[251f_{n+1} + 646f_n - 264f_{n-1}\right. \\ \qquad\qquad\qquad\qquad \left. +106f_{n-2} - 19f_{n-3}\right] \\ \tau(x,y;h) = -\dfrac{3h^5}{160}\,(d_f^5 f)(x,y) + \mathcal{O}(h^6) \end{cases}$$