Apontamentos de Matemática Computacional

Mário Meireles Graça
e
Pedro Trindade Lima

Departamento de Matemática
Instituto Superior Técnico
Universidade de Lisboa

Conteúdo

Prefácio

Estes *Apontamentos* destinam-se a servir de texto de apoio às aulas de Matemática Computacional, disciplina oferecida pelo Departamento de Matemática do Instituto Superior Técnico, nomeadamente ao segundo ano dos cursos de Engenharia de Materiais, Engenharia Geológica e de Minas, Engenharia Electrónica, Engenharia de Telecomunicações e Informática, Engenharia do Ambiente, Engenharia Química, Engenharia Civil e Engenharia Electrotécnica e de Computadores – para alunos do Campus da Alameda e do Taguspark – e alguns cursos da Academia da Força Aérea.

Depois da "Reforma de Bolonha", a disciplina de *Métodos Numéricos* foi substituída por *Matemática Computacional*. Com essa mudança desapareceram as aulas práticas e o tempo lectivo desta disciplina reduziu-se a três horas por semana (42 horas por semestre).

Segundo essa Reforma, uma missão dos alunos é aprender a estudar. Espera-se que o presente texto os possa ajudar.

Os Autores

Matérias não obrigatórias estão assinaladas no índice geral com asterisco (como por exemplo *4.2 Interpolação polinomial bivariada **). Além dos exercícios propostos ao longo dos capítulos, no Suplemento (A), página 329, são apresentados vários testes e exames resolvidos.

Os autores agradecem antecipadamente a todos os que desejem assinalar erros ou imperfeições, através dos endereços

mario.meireles.graca@tecnico.ulisboa.pt
ou
pedro.t.lima@tecnico.ulisboa.pt.

Instituto Superior Técnico, Universidade de Lisboa, Março de 2022.

Capítulo 1

Elementos da teoria dos erros computacionais

1.1 Representação de números. Erros de arredondamento.

1.1.1 Sistemas de ponto flutuante

Para efectuarmos cálculos é necessário antes de mais escolher um sistema de representação dos números. Supondo que vamos trabalhar com números reais, os sistemas habitualmente utilizados para os representar são designados por *sistemas de ponto flutuante* (ou de *vírgula flutuante*). Começamos por definir tais sistemas.

Seja $\beta \geq 2$ um número natural, a que chamaremos *base* do sistema. A base indica o número de dígitos distintos que usamos para representar os números. A base mais corrente é a decimal, $\beta = 10$, em que se usam dez dígitos (ou algarismos).

Um número real $x \neq 0$ pode ser representado numa dada base como $x = \pm\,(\text{parte inteira}) \cdot (\text{parte fraccionária})$,

$$x = \pm(a_n a_{n-1} \cdots a_1 a_0 \cdot a_{-1} a_{-2} \cdots a_{-m} \cdots),$$

onde os dígitos $a_i \in \{0, 1, \ldots, \beta - 1\}$. O valor de x é

$$\pm a_n \times \beta^n + a_{n-1} \times \beta^{n-1} + \ldots + a_1 \times \beta + a_0 + a_{-1} \times \beta^{-1} + a_{-2} \times \beta^{-2} + \ldots$$

Por exemplo, $\pi = 3.1415\cdots = 0.00031415\cdots \times 10^4 = 0.31415\cdots \times 10^1 = 31.415\cdots \times 10^{-1}$, ou qualquer outra representação onde se ajuste convenientemente o expoente da base 10. Para se evitar ambiguidade na representação, adopta-se a chamada *representação normalizada*,

$$x = \pm\,.(a_1 a_2 \cdots a_n \cdots) \times \beta^t, \qquad a_1 \geq 1, \quad t \in \mathbb{Z}\ .$$

Assim, um número x é representado na forma

$$x = \pm\, m \times \beta^t,$$

onde $0 < m < 1$ é habitualmente designado por *mantissa*, e t por *expoente*. A mantissa pode conter uma infinidade de dígitos, mas o seu primeiro dígito é sempre maior ou igual a 1.

Se atendermos à forma como os números são representados internamente nos computadores e noutros sistemas de cálculo, verificamos que a base aí utilizada é usualmente a binária, ou seja $\beta = 2$, já que por razões técnicas é conveniente trabalhar-se apenas com dois símbolos diferentes, 0 e 1. Nesse caso, cada símbolo representado designa-se por *bit.*

Uma vez escolhida a base, qualquer elemento do sistema de vírgula flutuante será denotado por $fl(x)$. Ao contrário dos números reais, cuja representação pode conter uma infinidade de dígitos, um número num sistema flutuante possui representação finita. Tal número assume a forma

$$fl(x) = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t, \tag{1.1}$$

onde σ representa o sinal ($\sigma = \pm 1$), os símbolos a_i representam dígitos na base considerada, e t é um número inteiro.

Admitimos que o número $fl(x)$ está escrito na *forma normalizada*, i.e., $a_1 \geq 1$. Assim, além da base, qualquer sistema de ponto flutuante caracteriza-se pelo *comprimento da mantissa,* isto é, o número n de dígitos que a compõem. Finalmente, um tal sistema depende ainda dos limites inferior e superior do expoente t, que representaremos respectivamente por t_1e t_2. Chegamos assim à seguinte definição.

Definição 1.1. (Sistema de ponto flutuante com base β e n dígitos na mantissa)

$$\begin{aligned} FP(\beta, n, t_1, t_2) \quad &= \{x \in \mathbb{R} : \quad x = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t, \\ &\quad \sigma = \pm 1, \quad a_1 \geq 1, \quad t_1 \leq t \leq t_2, \quad t \in \mathbb{Z}\} \cup \{0\} \ . \end{aligned}$$

Usamos a nomenclatura FP (de *floating-point*) ou VF (de *vírgula flutuante*) para indicar tratar-se de um sistema de representação de números como se descreveu anteriormente. De acordo com a Definição 1.1, como é natural, o número 0 pertence a qualquer sistema FP, embora formalmente ele não possa ser representado na forma (1.1), já que o primeiro dígito da mantissa de um número normalizado é diferente de zero. Daí que num sistema FP o número 0 tenha uma representação à parte.

Exemplo 1.1. *Considere uma calculadora em que os números são representados na base decimal, usando* 12 *dígitos na mantissa e expoente* t *entre* -99 *e* 99*. Como é representado o número* $x = 100$*, nesse sistema?*

O sistema utilizado é $FP(10, 12, -99, 99)$. O número 100 é representado como

$$+0.100000000000 \times 10^3.$$

♦

Exemplo 1.2. *Considere um computador em que os números são representados na base binária, sendo reservados 56 bits para a mantissa e 8 bits para o expoente. Suponha que 7 dos 8 bits do expoente são reservados ao seu valor absoluto e um ao sinal, pelo que o valor representado pelo expoente t pode variar entre* $-2^7+1 = -127$ *e* $2^7 - 1 = 127$. *Logo, o sistema considerado é* $VF(2, 56, -127, 127)$.

O número $x = 1/10$ *existe nesse sistema?*

O número em causa é representado na base 2 como $(0.1)_2 = 0.0001100110011\ldots$, ou seja possui um número infinito de bits que se repetem periodicamente, logo não existe em $VF(2, 56, -127, 127)$[1]. ♦

Note-se que, quando a base é $\beta = 2$, devido à condição $a_1 \geq 1$, no caso do sistema binário o primeiro dígito da mantissa é $a_1 = 1$, qualquer que seja o número não nulo representado. Isto faz com que esse dígito da mantissa seja supérfluo, e como tal é tomado como implícito na representação normalizada de números binários em computador.

1.1.2 Algumas propriedades dos sistemas de ponto flutuante

1. *Qualquer sistema VF é finito.*

Determinemos o número de elementos positivos do sistema $VF(\beta, n, t_1, t_2)$.
O número de mantissas diferentes é $\beta^{n-1}(\beta - 1)$ (o primeiro dígito da mantissa não pode ser 0). O número de expoentes diferentes é $t_2 - t_1 + 1$. Logo, o número N de elementos do sistema $VF(\beta, n, t_1, t_2)$, tendo em conta os números negativos e o zero, é

$$N = 2\,\beta^{n-1}(\beta - 1)\,(t_2 - t_1 + 1) + 1\ .$$

No caso do Exemplo 1.1, obtém-se $N = 2 \times 9 \times 10^9 \times 199 + 1 \approx 3.6 \times 10^{12}$ elementos, enquanto que para o Exemplo 1.2, o número de elementos é $N = 2 \times 255 \times 2^{55} + 1 \approx 1.84 \times 10^{19}$.

2. *Um sistema VF é limitado.*

3. *Um sistema* $FP(\beta, n, t_1, t_2)$ *contém apenas uma parte dos números racionais, isto é* $FP \subset \mathbb{Q}$.

[1] Deixa-se ao leitor a tarefa de confirmar se a representação binária de $(0.1)_{10}$ é a que se refere.

De facto, sendo $fl(x) > 0 \in FP$, tal que $fl(x) = (0.a_1\, a_2, \ldots\, a_n)_\beta \times \beta^t$, o número é racional sendo o seu valor

$$(a_1 \times \beta^{-1} + a_2 \times \beta^{-2} + \ldots + a_n \times \beta^{-n}) \times \beta^t, \qquad \in \mathbb{Q}\ .$$

Se M e m representarem respectivamente o maior e o menor elemento positivo do sistema, tem-se

$$M = (1 - \beta^{-n})\beta^{t_2}$$
$$m = \beta^{-1}\beta^{t_1} = \beta^{t_1 - 1}.$$

No caso do Exemplo 1.1, obtém-se $M = (1 - 10^{-12})10^{99} \approx 10^{99}$ e $m = 10^{-100}$, enquanto que para o Exemplo 1.2 é $M = (1 - 2^{-155})2^{127} \approx 1.70 \times 10^{38}$ e $m = 2^{-128} \approx 2.9 \times 10^{-39}$.

A implementação em computador de um sistema de representação numérica normalizada obedece a regras definidas pelo *Institute of Electrical and Electronics Engineers* ($IEEE$).

A tabela a seguir indica alguns parâmetros adoptados nos sistemas FP usuais, para $\beta = 2$, segundo a norma IEEE754–2008[2] desse organismo.

	n	t_1	t_2
binary32	24	-125	128
binary64	53	-1021	1024
binary128	113	$-16\,381$	$16\,384$

1.1.3 Arredondamentos

Tal como se disse anteriormente, qualquer sistema FP contém uma parte dos números reais constituida apenas por um número finito de números racionais.

Quando um número real não pertence ao sistema considerado, para o representar nesse sistema é necessário fazer uma certa aproximação, chamada *arredondamento*. Basta lembrar-se do que acontece ao representar $(0.1)_{10}$ na base 2.

Denotemos por $fl(x)$ a representação do número real $x > 0$ no sistema VF considerado. Se $x \in VF(\beta, n, t_1, t_2)$, então $fl(x) = x$ (diz-se que x tem representação exacta nesse sistema). Caso contrário, isto é, se $x \notin VF(\beta, n, t_1, t_2)$, mas $m \leq x \leq M$, há que atribuir um valor do sistema FP a $fl(x)$, e essa escolha pode ser feita de diferentes maneiras.

Para melhor compreender este processo, suponhamos que

$$x = \sigma \times 0.a_1a_2a_3 \cdots a_na_{n+1} \cdots \times \beta^t.$$

[2] http://standards.ieee.org/.

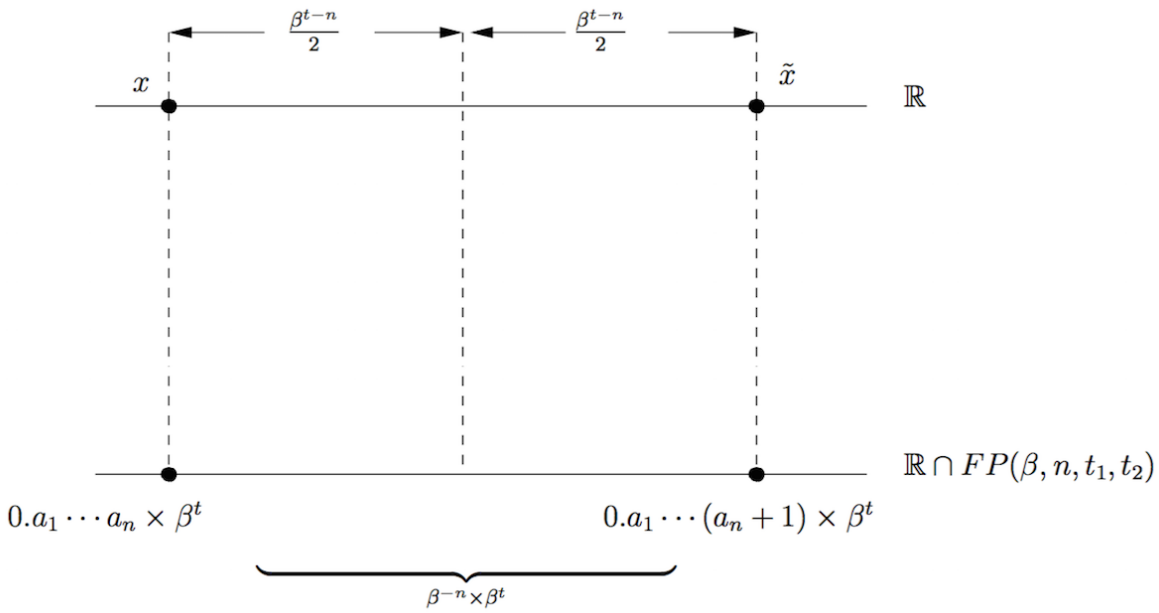


Figura 1.1: Distância entre números consecutivos de $FP(\beta, n, t_1, t_2)$.

Relembre-se que qualquer número real não nulo pode ser representado nesta forma, sendo que a mantissa, regra geral, é infinita. Segundo a forma mais simples de arredondamento, o *arredondamento por corte*, resulta

$$fl(x) = \sigma \times 0.a_1 a_2 a_3 \cdots a_n \times \beta^t.$$

Outra forma de obter $fl(x)$ consiste em defini-lo através de

$$fl(x) = \begin{cases} \sigma \times 0.a_1 a_2 a_3 \cdots a_n \times \beta^t, \text{ se } a_{n+1} < \beta/2 \\ \sigma \times ((0.a_1 a_2 a_3 \cdots a_n) + \beta^{-n}) \times \beta^t \Big\|_{\text{(normalizar)}}, \text{ se } a_{n+1} \geq \beta/2, \end{cases} \tag{1.2}$$

o que corresponde à noção habitual de arredondamento de números. Esta forma de aproximação chama-se *arredondamento simétrico.*

O arredondamento simétrico envolve um erro igual ao do arredondamento por corte, no caso de $a_{n+1} < \beta/2$, ou menor, no caso em que $a_{n+1} \geq \beta/2$ (ver Fig. 1.1).

"Overflow/underflow"

Ao considerar um certo sistema $FP(\beta, n, t_1, t_2)$, há números reais que não podem ser representados. Os números x, tais que $|x| > M$ ou $|x| < m$, não têm

representação no sistema, pelo que ocorrem situações de erro ao tentar representá-los. No primeiro caso, essas situações designam-se por *overflow*, enquanto no segundo caso são referidas como *underflow*. Os construtores de máquinas de cálculo adoptam estratégias de aviso de ocorrência dessas situações através de mensagens apropriadas.

Exemplo 1.3. *Para cada um dos seguintes números reais obter (caso seja possível) a sua representação no sistema* $VF(10, 3, -99, 99)$*, utilizando arredondamento simétrico.*

a) $x = 100$*;*
b) $x = 0.001235$*;*
c) $x = -1001$*;*
d) $x = 1/3$*;*
e) $x = 10^{100}$*;*
f) $x = 10^{-101}$*;*
g) $x = 9.999$*.*

Na tabela a seguir apresentamos as respostas às alíneas anteriores.

x	$fl(x)$
100	0.100×10^3
0.001235	0.124×10^{-2}
-1001	-0.100×10^4
1/3	0.333
10^{100}	não tem representação (overflow)
10^{-101}	não tem representação (underflow)
9.999	0.100×10^2

Note que na alínea g) o número $fl(x)$ resultou de adicionar 0.01 a 9.99 e normalizar o resultado. Todos os dígitos da mantissa de $fl(x)$ diferem dos dígitos do valor inicial x (esta é a razão que justifica que em (1.2) se faça uma normalização do resultado). ♦

1.1.4 Erros de arredondamento

Quando se aproxima um número real x pela sua representação em ponto flutuante, $fl(x)$, comete-se em geral um erro designado por *erro de arredondamento*,

$$e_{ar} = fl(x) - x \ .$$

Grandezas relacionadas com e_{ar} são: o *erro de arredondamento absoluto*

$$|e_{ar}| = |x - fl(x)| \, ,$$

e o *erro de arredondamento relativo*,

$$|\delta_{ar}| = \frac{|x - fl(x)|}{|x|}, \qquad x \neq 0 \ .$$

Para caracterizarmos um sistema $FP(\beta, n, t_1, t_2)$ em termos duma estimativa da grandeza dos erros previsíveis de arredondamento, consideremos um número real x arbitrário e representemo-lo na forma normalizada

$$x = \sigma \times 0.a_1 a_2 a_3 \cdots a_n a_{n+1} \cdots \times \beta^t, \qquad a_1 \geq 1 \ .$$

Na Figura 1.1, pág. 7, está representado o segmento de números reais entre $x > 0$ e o número $\tilde{x}$, cujo último dígito da mantissa difere de uma unidade do n-ésimo dígito da mantissa de x. Os números x e $\tilde{x}$ possuem representação exacta no sistema em causa e são, portanto, dois números consecutivos deste sistema. A distância entre esses números consecutivos vale β^{t-n}. Qualquer número real do segmento $[x, \tilde{x}]$ será representado no sistema FP ou por $0.a_1 \cdots a_n \times \beta^t$, ou por $0.a_1 \cdots (a_n + 1) \times \beta^t$.

Comecemos por considerar o caso do *arredondamento por corte*. Como já vimos, neste caso $fl(x) = \sigma \times 0.a_1 a_2 a_3 ... a_n \times \beta^t$. Por conseguinte, o erro de arredondamento absoluto satisfaz a desigualdade

$$|e_{ar}| = |x - fl(x)| = 0.00...0a_{n+1}... \times \beta^t < \beta^{t-n}.$$

No que diz respeito ao erro de arredondamento relativo, temos

$$|\delta_{ar}| = \frac{|x - fl(x)|}{|x|} \leq \frac{|x - fl(x)|}{(0.10 \cdots 0)_\beta \times \beta^t} < \frac{\beta^{t-n}}{\beta^{t-1}} = \beta^{1-n}.$$

Assim, qualquer que seja x, tal que $m \leq |x| \leq M$, verifica-se

$$|\delta_{ar}| < \beta^{1-n}. \tag{1.3}$$

Unidade de arredondamento

Para caracterizar a precisão com que os números reais são aproximados num sistema FP utiliza-se o conceito de *unidade de arredondamento*.

Definição 1.2. A unidade de arredondamento de um sistema $FP(\beta, n, t_1, t_2)$ é um número real u, tal que

$$|\delta_{ar}| \leq u, \qquad \forall x \in \mathbb{R}, \quad m \leq |x| \leq M.$$

A unidade de arredondamento u é por conseguinte um majorante do erro relativo máximo de arredondamento quando se passa de x a $fl(x)$.

O valor de u depende, evidentemente, dos parâmetros do sistema considerado, mais precisamente, de n e β. Para o mesmo valor da base β, a unidade de arredondamento será tanto mais pequena quanto maior for n, isto é, quanto mais dígitos utilizarmos para representar os números tanto menor será o erro de arredondamento relativo.
Logo, de (1.3), resulta que, no caso do arredondamento por corte, a unidade de arredondamento é[3]

$$u = \beta^{1-n}. \tag{1.4}$$

Levando em consideração a Figura 1.1, pág. 7, e o que se disse sobre o modo de obtenção de um número por *arredondamento simétrico*, neste caso a respectiva unidade de arredondamento é

$$u = \frac{1}{2}\beta^{1-n}. \tag{1.5}$$

Por exemplo, no caso do sistema $VF(10, 12, -99, 99)$, e assumindo que o arredondamento é simétrico, temos $u = 0.5 \times 10^{-11}$.

Exemplo 1.4. *(a) Considerando de novo o sistema* $VF(10, 3, -99, 99)$*, para cada um dos números reais* x *da tabela abaixo, estão calculados o erro de arredondamento absoluto e o erro de arredondamento relativo. Compare este último com a unidade de arredondamento simétrico do sistema.*

(b) Qual é a distância entre o número 1 *e o número imediatamente superior a* 1 *representado no sistema? Há alguma relação entre essa distância e a unidade de arredondamento* u*?*

(a)

x	$fl(x)$	$\lvert e_{ar}\rvert$	$\lvert\delta_{ar}\rvert$
100	0.100×10^{3}	0	0
0.001235	0.124×10^{-2}	0.5×10^{-5}	0.004
-1001	-0.100×10^{4}	1	0.001
$1/3$	0.333	0.33×10^{-3}	0.001
0.9995	0.100×10^{1}	0.5×10^{-3}	0.5002×10^{-3}

A unidade de arredondamento simétrico vale $0.5 \times 10^{-2} = 0.005$, pelo que todos os números considerados possuem erro de arredondamento relativo inferior a u.

(b) Como $1 = 0.100 \times 10^{1}$ e o número representado imediatamente superior é $\tilde{1} = 0.101 \times 10^{1}$, a referida distância é $0.001 \times 10^{1} = 10^{1-3} = 2\,u$, ou seja, vale o dobro da unidade de arredondamento.

[3]Expressões como (1.4) estão reunidas num Formulário, pág. 329.

De modo análogo, a distância entre 10 e o número representado imediatamente superior, seja $\bar{10}$, passa a ser $2\,u * 10$. Tal significa que a unidade de arredondamento mede a "granularidade"do sistema. Com efeito, dois números consecutivos representados no sistema encontram-se cada vez mais afastados entre si à medida que a ordem de grandeza (dada pelo expoente t) aumenta. Apesar disso, na passagem de x a $fl(x)$, o erro relativo que se comete nunca é superior à unidade de arredondamento, independentemente da grandeza de x. ♦

1.1.5 Propagação dos erros

Sejam $\bar{x}$ e $\bar{y}$ valores aproximados dos números reais x e y, respectivamente. Denotaremos por $|e_{\bar{x}}|$ e $|\delta_{\bar{x}}|$ respectivamente os erros absoluto e relativo de $\bar{x}$,

$$e_{\bar{x}} = \bar{x} - x,$$

$$|\delta_{\bar{x}}| = \left|\frac{\bar{x} - x}{x}\right|, \quad x \neq 0 \ .$$

De modo análogo se definem os erros de $\bar{y}$. Suponhamos que $\bar{x}$ e $\bar{y}$ são dados de um cálculo que pretendemos efectuar. O nosso objectivo é determinar qual o efeito dos erros dos dados no resultado. Para começar, consideremos o caso das operações aritméticas.

Adição/Subtracção

Representemos por $e_{\bar{x}\pm\bar{y}}$o erro de $\bar{x} \pm \bar{y}$. Note-se que

$$\bar{x} \pm \bar{y} = (x + e_{\bar{x}}) \pm (y + e_{\bar{y}}) = (x \pm y) + (e_{\bar{x}} \pm e_{\bar{y}}) \ .$$

Por conseguinte, para o erro de $\bar{x} \pm \bar{y}$ temos

$$e_{\bar{x}\pm\bar{y}} = e_{\bar{x}} \pm e_{\bar{y}}$$

e, para o erro absoluto,

$$|e_{\bar{x}\pm\bar{y}}| \leq |e_{\bar{x}}| + |e_{\bar{y}}| \ .$$

Quanto ao erro relativo, podemos escrever

$$|\delta_{\bar{x}\pm\bar{y}}| = \frac{|e_{\bar{x}} \pm e_{\bar{y}}|}{|x \pm y|} \leq \frac{|x\,\delta_{\bar{x}}| + |y\,\delta_{\bar{y}}|}{|x \pm y|} \ . \tag{1.6}$$

Daqui resulta que, se o valor de $x \pm y$ for próximo de zero, então o erro relativo do resultado pode ser muito maior que o dos dados $\bar{x}$ e $\bar{y}$. Voltaremos a este assunto adiante, (ver pág. 16).

Multiplicação

No caso da multiplicação, temos

$$\bar{x}\,\bar{y} = (x + e_{\bar{x}}) \times (y + e_{\bar{y}}) = x\,y + y\,e_{\bar{x}} + x\,e_{\bar{y}} + e_{\bar{x}}\,e_{\bar{y}} \ .$$

Admitindo que $|e_{\bar{x}}|$ e $|e_{\bar{y}}|$ são grandezas pequenas, o seu produto pode ser desprezado na expressão anterior, pelo que obtemos

$$e_{\bar{x}\times\bar{y}} = \bar{x} \times \bar{y} - x \times y \approx y\,e_{\bar{x}} + x\,e_{\bar{y}} \ .$$

Logo, para o erro relativo do produto resulta

$$|\delta_{\bar{x}\times\bar{y}}| = \frac{|e_{\bar{x}\times\bar{y}}|}{|x \times y|} \approx \frac{|y\,e_{\bar{x}} + x\,e_{\bar{y}}|}{|x \times y|} \leq |\delta_{\bar{x}}| + |\delta_{\bar{y}}| \ . \tag{1.7}$$

Divisão

Para deduzir uma aproximação do erro do quociente, suponhamos que os valores de $|e_{\bar{x}}|$ e $|e_{\bar{y}}|$ são desprezáveis em comparação com $|x|$ e $|y|$, respectivamente. Podemos então fazer a seguinte aproximação,

$$\frac{\bar{x}}{\bar{y}} = (x + e_{\bar{x}})\,\frac{1}{y}\,\frac{1}{1 + \dfrac{e_{\bar{y}}}{y}} \approx (x + e_{\bar{x}})\,\frac{1}{y}\left(1 - \frac{e_{\bar{y}}}{y}\right) \approx \frac{x}{y} + \frac{y\,e_{\bar{x}} - x\,e_{\bar{y}}}{y^2},$$

donde

$$e_{\bar{x}/\bar{y}} = \frac{\bar{x}}{\bar{y}} - \frac{x}{y} \approx \frac{y\,e_{\bar{x}} - x\,e_{\bar{y}}}{y^2} \ .$$

Quanto ao erro relativo do quociente, obtém-se

$$|\delta_{\bar{x}/\bar{y}}| = |e_{\bar{x}/\bar{y}}|\frac{|y|}{|x|} \approx \frac{|y\,e_{\bar{x}} - x\,e_{\bar{y}}|}{y^2}\frac{|y|}{|x|} \leq \frac{|e_{\bar{x}}|}{|x|} + \frac{|e_{\bar{y}}|}{|y|} = |\delta_{\bar{x}}| + |\delta_{\bar{y}}| \ . \tag{1.8}$$

Com rigor as majorações dadas pelas expressões (1.7) e (1.8) não são propriamente majorações, mas antes aproximações de majorações. Essas expressões servirão todavia como modelo de propagação de erro, permitindo efectuar *estimativas* de erro.

Cancelamento subtractivo

Os cálculos anteriores mostram que, no caso da multiplicação e da divisão, o erro relativo dos resultados é da mesma ordem de grandeza que o erro relativo dos dados, ou seja, destas operações não resulta uma perda de precisão. Já no caso da adição e da subtracção, como vimos, tal perda de precisão pode ocorrer. Esse fenómeno designa-se por *cancelamento subtractivo*. Uma ilustração é dada no Exemplo 1.5, pág. 13.

As estimativas de erro que fizemos para as operações binárias $+$, $-$, $\times$ e :, poderão ser obtidas mais facilmente usando estimativas de erro propagado por funções (ver secção 1.2, pág. 16).

Exemplo 1.5. *Considere os números $x = \pi$ e $y = 2199/700$.*

(a) Determine aproximações $\bar{x}$ e $\bar{y}$ com 4 dígitos na mantissa, usando arredondamento simétrico. Obtenha ainda $\bar{x} - \bar{y}$.

(b) Calcule os erros absolutos e relativos de $\bar{x}$ e $\bar{y}$. Comente.

(c) Represente os números x e y em ponto flutuante, mas com 6 algarismos na mantissa. Com base nestas novas aproximações, calcule de novo $\bar{x} - \bar{y}$ e comente.

(d) Tomando como valor exacto da diferença o resultado da alínea anterior, determine o erro relativo do valor de $\bar{x} - \bar{y}$, obtido na alínea (a). Se usasse a estimativa (1.6) para o erro relativo da diferença, chegaria à mesma conclusão?

(a)

$$\begin{array}{ll} x = 0.3141592\cdots \times 10^1, & \bar{x} = fl(x) = 0.3142 \times 10^1 \\ y = 0.3141428\cdots \times 10^1, & \bar{y} = fl(y) = 0.3141 \times 10^1. \end{array}$$

Logo, $\bar{z} = \bar{x} - \bar{y} = 0.1 \times 10^{-2}$.

(b)

Dado	Erro absoluto	Erro relativo
x	0.41×10^{-3}	0.131×10^{-3}
y	0.43×10^{-3}	0.137×10^{-3}

Como seria de esperar, os erros de arredondamento relativos dos dados são inferiores à unidade de arredondamento simétrico do sistema que, neste caso, é $u = 0.5 \times 10^{1-4} = 0.5 \times 10^{-3}$.

(c) Neste caso temos:

$$\begin{array}{ll} x = 0.3141592\cdots \times 10^1 & \bar{x} = fl(x) = 0.314159 \times 10^1 \\ y = 0.3141428\cdots \times 10^1 & \bar{y} = fl(y) = 0.314143 \times 10^1. \end{array}$$

Logo, $\tilde{z} = \tilde{x} - \tilde{y} = 0.16 \times 10^{-3}$, o que é um valor cerca de 10 vezes menor do que o obtido na alínea (a). Isto sugere que, na alínea (a), houve uma perda de precisão resultante de cancelamento subtractivo.

(d) Comparando os resultados das alíneas (a) e (c), para $\bar{z} = \bar{x} - \bar{y}$, temos

$$|\delta_{\bar{x}-\bar{y}}| = \frac{|e_{\bar{x}-\bar{y}}|}{|x-y|} \approx \frac{0.001 - 0.00016}{0.00016} = 5.25 = 525\,\% \ .$$

Vemos que o erro relativo do resultado $\bar{z}$ da alínea (a) é muito superior à unidade, o que significa uma perda total de precisão. ♦

1.1.6 Estabilidade de algoritmos

Quando se efectua um cálculo, geralmente ele é processado passo a passo. Assim, o erro cometido em cada passo acumula-se eventualmente com os erros cometidos nos passos anteriores. Por conseguinte, o erro do resultado final pode ser muito maior do que o erro cometido isoladamente em cada passo.

Por exemplo, vamos assumir que a tarefa de calcular o valor de uma determinada expressão algébrica foi fragmentada através de "operações elementares", como sejam dividir por um número, somar, subtrair, multiplicar ou dividir dois números ou, por exemplo, calcular $\sqrt{(.)}$, $\sin(.)$, ou um valor exponencial $e^{\cdot}$, onde o símbolo "(.)" representa um certo argumento. De modo informal, dizemos que um procedimento sistemático com vista à obtenção de um dado resultado é um *algoritmo*. Assim, consideraremos um algoritmo como sendo um conjunto ordenado de tarefas elementares, ou *passos*.

Em particular, o resultado de um determinado cálculo pode ser obtido, em princípio, através de algoritmos distintos. No entanto, os erros propagam-se de forma diferente em cada algoritmo, visto que ao executarmos sequências distintas de operações elementares estaremos a cometer erros dependentes dessas operações. Por isso, os resultados que se obtêm para o mesmo problema, através de algoritmos distintos, podem possuir precisões significativamente diferentes. Surge assim a definição de *estabilidade numérica*.

Definição 1.3. Um algoritmo diz-se estável (ou numericamente estável) para um certo conjunto de dados se, a pequenos valores dos erros relativos de arredondamento dos dados (e da unidade de arredondamento do sistema) corresponderem pequenos valores do erro relativo do resultado.

O Exemplo 1.5 ilustra o conceito de estabilidade numérica.

Exemplo 1.6. *Considere a função real de variável real*

$$f(x) = \frac{1 - \cos(x)}{x^2}, \qquad x > 0 \tag{1.9}$$

(a) Supondo que utiliza um sistema de vírgula flutuante com 10 *dígitos na mantissa e arredondamento simétrico, calcule* $f(10^{-6})$ *aplicando a fórmula (1.9).*

(b) Obtenha uma aproximação de $f(10^{-6})$*, utilizando o desenvolvimento de* f *em série de Taylor*[4]*, em torno de* $x = 0$*.*

(c) Sabendo que $1 - \cos x = 2\,\sin^2(x/2)$*, calcule* $f(10^{-6})$ *utilizando uma nova fórmula para* f*.*

(d) Compare os valores obtidos nas alíneas anteriores, e classifique os respectivos algoritmos quanto à estabilidade.

[4] $f(x) = f(0) + f'(0)\,x + \dfrac{f^{(2)}(0)}{2!}\,x^2 + \dfrac{f^{(3)}(0)}{3!}\,x^3 + \cdots$.

(a) A expressão (1.9) pode ser fragmentada num algoritmo com 3 passos. O resultado (exacto) de cada operação elementar será designado por z_i, $i = 1:3$.[5] O resultado calculado em cada passo é denotado por $\bar{z}_i$. Sendo $x = 10^{-6}$, temos

$$\begin{array}{l|l} z_1 = \cos(x) = 1 & \bar{z}_1 = 1 \\ z_2 = 1 - z_1 & \bar{z}_2 = 0 \\ z_3 = \dfrac{z_2}{x^2} & \bar{z}_3 = 0 \ . \end{array} \tag{1.10}$$

Note que a função f é contínua para $x > 0$ e $\lim_{x\to 0^+} = 1/2$. Por conseguinte, o valor de $f(10^{-6})$ deverá ser próximo de 0.5, pelo que o valor calculado não faz nenhum sentido.

Coloca-se a questão de saber se há algo de "errado" com a função $f(x)$ dada. Veremos adiante, quando discutirmos o condicionamento de uma função real (ver parágrafo 1.2.2, pág. 20), que a função em causa não tem nada de suspeito. A disparidade entre o valor calculado para $f(10^{-6})$, e o valor exacto da função no ponto 10^{-6}, deve-se exclusivamente ao algoritmo que foi adoptado. Por exemplo, tal desconformidade entre o valor calculado e o valor esperado desaparece se considerarmos um desenvolvimento de Taylor da função, como se mostra a seguir.

(b) Como é sabido, para valores de x próximos de zero, a função $cos(x)$ admite o seguinte desenvolvimento em série de Taylor:

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} + O(x^6),$$

donde,

$$f(x) = \frac{1 - \cos(x)}{x^2} = \frac{1}{2} - \frac{x^2}{4!} + O(x^4) \ . \tag{1.11}$$

Desprezando o termo $O(x^4)$ na fórmula (1.11), num sistema VF com 10 dígitos, obtém-se $f(10^{-6}) = 0.5000000000$.

(c) Uma expressão equivalente a (1.9) é

$$f(x) = \frac{1 - \cos(x)}{x^2} = \frac{2}{x^2}\,\sin^2(x/2) \ . \tag{1.12}$$

Apliquemos a expressão mais à direita em (1.12), considerando o seguinte algoritmo em 5 passos,

$$\begin{array}{l|l} w_1 = x/2 & \bar{w}_1 = 0.5 \times 10^{-6} \\ w_2 = \sin(w_1) & \bar{w}_2 = 0.5 \times 10^{-6} \\ w_3 = w_2^2 & \bar{w}_3 = 0.25 \times 10^{-12} \\ w_4 = w_3/x^2 & \bar{w}_4 = 0.25 \\ w_5 = f(x) = 2 \times w_4 & \bar{w}_5 = 0.5 \ . \end{array} \tag{1.13}$$

[5] Uma expressão como $i = m:n$, com $m < n$, significa $i = m, m+1, \ldots, n-1, n$.

(d) Verifica-se que o valor obtido em (c) é uma boa aproximação de $f(10^{-6})$, já que coincide com o valor dado pela série de Taylor e é próximo de $1/2$, como seria de esperar. Pelo contrário, o valor obtido pelo algoritmo da alínea (a) é uma má aproximação (que não possui sequer um único dígito correcto). Este facto deve-se apenas aos (pequenos) erros de arredondamento cometidos em cada passo, os quais aparecem muito ampliados no resultado final.

Os resultados obtidos podem interpretar-se do seguinte modo: para valores de x próximos de zero o algoritmo considerado em (a) é instável, enquanto o algoritmo considerado em (c) é estável. Na secção 1.3, pág. 23, discutiremos mais detalhadamente o conceito de *estabilidade* ou *instabilidade* numérica de um algoritmo.

♦

1.2 Fórmulas diferenciais de propagação de erro

A propagação de erros de arredondamento nas operações binárias de adição, subtracção, multiplicação e divisão, tratadas no parágrafo 1.1.5, pág. 11, usando as definições de erro absoluto e relativo, pode encarar-se como um caso particular da propagação de erro (quer seja de arredondamento ou não) a uma função real multivariada, quando se cometem erros nas variáveis independentes da função.

Esta abordagem mais geral permite-nos lidar com a propagação de erro duma forma mais abrangente (de modo a tratar inclusive o caso da propagação de erro em algoritmos). Para esse efeito iremos deduzir algumas fórmulas de propagação de erro que serão aqui designadas por *fórmulas diferenciais de propagação de erro*.

Fixado o inteiro $n \geq 1$, considere-se uma função $f : D \subset \mathbb{R}^n \mapsto \mathbb{R}$, onde D é um domínio convexo. Sejam $x = (x_1, x_2, \ldots, x_n)$ e $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ dois vectores em D, e admitamos que se adoptou uma certa norma vectorial $||\cdot||$, em $\mathbb{R}^n$ (as normas vectoriais mais interessantes do ponto de vista computacional serão discutidas na secção 3.0.1, pág. 93).

Consideremos dois pontos x e $\bar{x}$ do domínio de f, suficientemente "próximos". Subentendemos que $\bar{x}$ é uma aproximação do vector exacto x, no sentido em que

$$e_{\bar{x}} = (e_{\bar{x}_1}, e_{\bar{x}_2}, \ldots, e_{\bar{x}_n}) = (x_1 - \bar{x}_1, x_2 - \bar{x}_2, \ldots, x_n - \bar{x}_n)$$

é tal que $||e_{\bar{x}}|| \leq \epsilon$, com $\epsilon < 1$.

Coloca-se a questão de saber se para um valor de ϵ pequeno, isto é, quando $\bar{x}$ está próximo de x, o erro na função $e_f = f(x) - f(\bar{x})$ também é (desejavelmente) pequeno.

Supondo f suficientemente diferenciável num aberto $A \subset D$, o desenvolvimento de Taylor da função f, em torno do ponto $\bar{x}$, escreve-se

$$\begin{aligned} f(x) &= f(\bar{x}) + f'(\bar{x}).(x-\bar{x}) + r(x) \\ &= f(\bar{x}) + f'_{x_1}(\bar{x})\, e_{\bar{x}_1} + f'_{x_2}(\bar{x})\, e_{\bar{x}_2} + \ldots + f'_{x_n}(\bar{x})\, e_{\bar{x}_n} + r(x), \end{aligned} \tag{1.14}$$

onde $f'_{x_i}(\bar{x}) = \dfrac{\partial f}{\partial x_i}(\bar{x})$, para $i = 1:n$ (o símbolo $i = 1:n$, significa que o índice i varia de 1 a n, ou seja, $i = 1, 2, \ldots, n$).

O termo $r(x)$ em (1.14) designa uma certa correcção, cujo módulo admitimos ser não superior ao produto de uma certa constante C, pelo quadrado do erro de $\bar{x}$ (em norma), isto é,

$$|r(x)| \leq C\, ||x - \bar{x}||^2,$$

onde C não depende de $\bar{x}$ nem de x. Neste caso podemos dizer que $|r(x)|$ é da ordem do quadrado da norma do erro em $\bar{x}$, o que se traduz na expressão

$$|e_f| = \mathcal{O}(||x - \bar{x}||^2)\ .$$

Fórmulas de propagação do erro absoluto

Supondo que $||x - \bar{x}||^2 << ||x - \bar{x}||$, podemos desprezar a correcção $r(x)$ em (1.14), obtendo-se a seguinte aproximação para o erro de $f(\bar{x})$,

$$e_{f(\bar{x})} = f(x) - f(\bar{x}) \ \simeq f'_{x_1}(\bar{x})\, e_{\bar{x}_1} + f'_{x_2}(\bar{x})\, e_{\bar{x}_2} + \ldots + f'_{x_n}(\bar{x})\, e_{\bar{x}_n}. \tag{1.15}$$

Como por hipótese, f' é contínua e $\bar{x}$ é próximo de x, é verdade que

$$\frac{\partial f}{\partial x_i}(\bar{x}) \simeq \frac{\partial f}{\partial\, x_i}(x), \qquad i = 1:n$$

pelo que, aplicando a fórmula de Taylor, podemos considerar a fórmula de aproximação do erro

$$e_{f(\bar{x})} = f(x) - f(\bar{x}) \simeq f'_{x_1}(x)\, e_{\bar{x}_1} + f'_{x_2}(x)\, e_{\bar{x}_2} + \ldots + f'_{x_n}(x)\, e_{\bar{x}_n}\ . \tag{1.16}$$

As fórmulas (1.15) e (1.16), embora sejam utilizadas adiante para finalidades distintas, recebem a designação de *fórmulas de propagação do erro absoluto*.

Atendendo à desigualdade triangular para o módulo, de (1.15) e (1.16) resultam as seguintes majorações do erro absoluto[6],

$$|e_{f(\bar{x})}| \leq |f'_{x_1}(\bar{x})|\, |e_{\bar{x}_1}| + |f'_{x_2}(\bar{x})|\, |e_{\bar{x}_2}| + \ldots + |f'_{x_n}(\bar{x})|\, |e_{\bar{x}_n}| \tag{1.17}$$

[6]Tal como já foi observado antes, trata-se de fórmulas aproximadas que servirão como estimativas do majorante de erro em causa.

e

$$|e_{f(\bar{x})}| \leq |f'_{x_1}(x)|\,|e_{\bar{x}_1}| + |f'_{x_2}(x)|\,|e_{\bar{x}_2|} + \ldots + |f'_{x_n}(x)|\,|e_{\bar{x}_n}| \ . \tag{1.18}$$

As duas fórmulas anteriores podem usar-se sempre que conhecermos majorações dos erros absolutos de cada uma das variáveis da função.

Exemplo 1.7. *Sabendo que o valor* 1.21 *resulta de um arredondamento simétrico, estimar o valor de* $\tan(1.21)$*, e concluir a respeito de quantos algarismos significativos se podem garantir para o valor estimado.*

Sejam x e $f(x) = \tan(x)$, valores exactos que desconhecemos. Sabemos apenas que $\bar{x} = 1.21$ e (usando uma máquina de calcular) que $f(\bar{x}) = \tan(\bar{x}) = 2.6503\cdots$. Uma vez que $\bar{x}$ resultou de um arredondamento simétrico, sabemos também que

$$|e_{\bar{x}}| = |x - \bar{x}| \leq \epsilon, \quad \text{com} \quad \epsilon = 0.5 \times 10^{-2}.$$

Dado que $f'(x) = \sec^2(x)$, de (1.17) obtém-se

$$|e_{\tan(\bar{x})}| \leq |f'(\bar{x})|\,|e_{\bar{x}}| \leq |f'(\bar{x})| \times \epsilon,$$

isto é,

$$|e_{\tan(\bar{x})}| \leq \sec^2(\bar{x}) \times 0.5 \times 10^{-2} \simeq 0.04012 \ .$$

Visto que o valor calculado $\tan(\bar{x}) = 2.6503\cdots$ possui um erro estimado que afecta a sua segunda casa decimal em cerca de 4 unidades dessa posição, concluímos intuitivamente que apenas os dois primeiros dígitos da aproximação deverão ser considerados significativos. Por conseguinte, será boa prática apresentar o resultado na forma

$$\tan(1.21) = 2.65 \pm 0.04,$$

dando assim uma indicação da "qualidade"da aproximação calculada. ♦

Número de algarismos significativos

O Exemplo 1.7 sugere a necessidade de se definir o conceito de *número de algarismos significativos* de uma aproximação, definição essa que seja coerente com a mesma noção intuitiva quando comparamos dois números, representados na base $\beta = 10$, em que um deles é considerado aproximação do outro. Por exemplo, pode dizer-se que o número $\bar{x} = 22/7 = 3.1428\cdots$ é uma aproximação com três algarismos significativos do número $x = \pi = 3.141592\cdots$, porquanto o erro absoluto de $\bar{x}$ manifesta-se apenas a partir da terceira casa decimal de x.

Assim, admitimos que se conhece a ordem de grandeza de um valor exacto x, através do expoente t da forma decimal normalizada desse valor. Ou seja,

$$|x| = 0.a_1\cdots \times 10^t, \qquad a_1 \geq 1 \ .$$

Sendo $\bar{x}$ uma aproximação de x, diremos que $\bar{x}$ possui um certo número k de algarismos significativos, se o seu erro absoluto não exceder meia unidade da k-ésima posição da mantissa de x, isto é,

Definição 1.4. Um número $|\bar{x}|$, aproximação do número decimal normalizado $|x| = 0.a_1a_2\cdots \times 10^t$, possui k $(k \geq 0)$ algarismos significativos se

$$0.5 \times 10^{t-(k+1)} \leq |x - \bar{x}| \leq 0.5 \times 10^{t-k}$$

No Exemplo 1.7, pág. 18, o valor de uma função é tal que $f(x) = 2.6\cdots = 0.26\cdots \times 10^1$, isto é, sabemos que a respectiva ordem de grandeza é dada por $t = 1$, e que $|f(x) - f(\bar{x})| \simeq 0.04$. Atendendo a que

$$0.005 < |e_{f(\bar{x})}| = 0.04 < 0.05 = 0.5 \times 10^{-1} = 0.5 \times 10^{1-2},$$

segundo a Definição 1.4, o número $f(\bar{x}) = 2.6503\cdots$ possui apenas 2 algarismos significativos.

1.2.1 Fórmulas de propagação do erro relativo

A qualidade de uma aproximação $\bar{f} = f(\bar{x})$, relativamente à quantidade exacta $f = f(x)$, é melhor traduzida através do erro relativo do que mediante o erro absoluto, como se observou no Exemplo 1.5, pág. 13.

Atendendo a que para $x \neq (0, 0, \ldots, 0)$ e $f(x) \neq 0$, se tem

$$\frac{\frac{\partial f}{\partial x_i}(\bar{x})\, e_{x_i}}{f(x)} = \frac{x_i \frac{\partial f(\bar{x})}{\partial x_i} \frac{e_{x_i}}{x_i}}{f(x)} = \frac{x_i \frac{\partial f(\bar{x})}{\partial x_i}\, \delta_{\bar{x}_i}}{f(x)},$$

de (1.15) e (1.16), podemos dizer que o erro relativo de $f(\bar{x})$ satisfaz as seguintes relações, ditas *fórmulas de propagação do erro relativo*:

$$\delta_{f(\bar{x})} \simeq \frac{x_1\, f'_{x_1}(\bar{x})}{f(x)}\, \delta_{\bar{x}_1} + \frac{x_2\, f'_{x_2}(\bar{x})}{f(x)}\, \delta_{\bar{x}_2} + \ldots + \frac{x_n\, f'_{x_n}(\bar{x})}{f(x)}\, \delta_{\bar{x}_n}. \tag{1.19}$$

$$\delta_{f(\bar{x})} \simeq \frac{x_1\, f'_{x_1}(x)}{f(x)}\, \delta_{\bar{x}_1} + \frac{x_2\, f'_{x_2}(x)}{f(x)}\, \delta_{\bar{x}_2} + \ldots + \frac{x_n\, f'_{x_n}(x)}{f(x)}\, \delta_{\bar{x}_n}. \tag{1.20}$$

A fórmula (1.19) é útil se se conhece o ponto aproximado $\bar{x}$ do ponto exacto x (geralmente desconhecido), ou seja, quando $f(\bar{x})$ é conhecido mas $f(x)$ não o é, havendo no entanto informação disponível a respeito do erro de $\bar{x}$.

Por sua vez, a fórmula (1.20) pode ser usada para prever o comportamento do erro da função f quando o vector argumento x (e consequentemente uma sua aproximação $\bar{x}$) percorre um certo domínio em $\mathbb{R}^n$, ou seja, para um certo conjunto de dados. Trata-se do estudo do chamado *condicionamento da função* f, que discutiremos na secção seguinte.

1.2.2 Condicionamento de uma função

A aproximação (1.20) mostra-nos que podem existir valores de alguma variável x_i da função f, para a qual a i-ésima parcela da referida fórmula de propagação de erro possua uma grandeza elevada, isto é, que a quantidade a seguir denotada por $P_{f,i}(x)$,

$$P_{f,i}(x) = \frac{x_i\, f'_{x_i}(x)}{f(x)}$$

seja tal que $|P_{f,i}(x)| >> 1$ (o símbolo $>>$ significa "muito maior", sendo que esse qualificativo estará dependente das quantidades em jogo em cada caso particular). A quantidade anterior é por vezes designada como o *peso* da função f relativamente à variável x_i.

Assim, quando $|P_{f,i}(x)|$ é grande, pode suceder que embora $|\delta_{\bar{x}_i}|$ seja pequeno, o correspondente termo em (1.20) possua um valor elevado. Quer isso dizer que o erro relativo propagado à função, $|\delta_{f(\bar{x})}|$, pode ser grande apesar de todos os erros relativos dos argumentos da função, $|\delta_{\bar{x}_i}|$, serem pequenos. Neste caso dizemos que a função f é *mal condicionada* para certo conjunto de dados $x = (x_1, x_2, \ldots, x_n)$, onde essa disparidade de grandezas de erros relativos se verifica. Tal justifica que os pesos em causa recebam a designação dada na seguinte definição.

Definição 1.5. O número

$$cond_{f,i}(x) = |P_{f,i}(x)| = \frac{|x_i\, f'_{x_i}(x)|}{|f(x)|}$$

diz-se número de condição de f relativamente à variável x_i, para $i = 1 : n$.
No caso de funções de uma só variável, o respectivo número de condição é simplesmente designado por $cond_f(x)$. A função associada à função f, definida pela expressão

$$cond_f(x) = \frac{|x\, f'(x)|}{|f(x)|}$$

diz-se função número de condição de f.

Exemplo 1.8. *Seja $a \in \mathbb{R}$ e*

$$f(x) = \frac{2}{x-a}, \qquad x \neq a \quad e \quad a \neq 0\ .$$

Fazendo, por exemplo, $a = 10$, a Fig. 1.2 mostra o gráfico da função $cond_f(x)$, para $0 \leq x \leq 60$. Dado que

$$\lim_{x \to 10} cond_f(x) = +\infty,$$

a função dada é mal condicionada para valores de x próximos de $a = 10$.

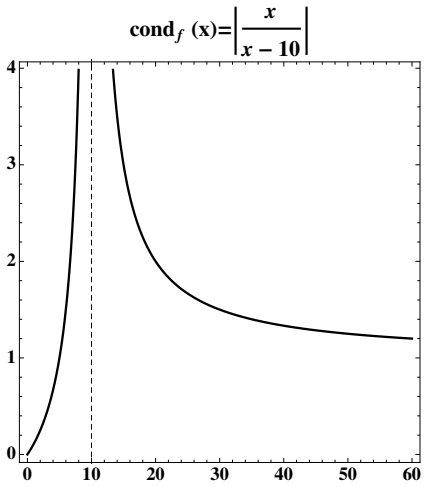


Figura 1.2: Função de condição para $f(x) = 2/(x-10)$

O mau condicionamento desta função para valores do denominador próximos de zero, explica por que razão geralmente se deve evitar a divisão de uma constante não nula por números próximos de zero.

Como ilustração, suponhamos que

$$x = a - \epsilon = 10 - \epsilon \quad \text{e} \quad \bar{x} = a + 2\,\epsilon, \quad \text{com} \quad \epsilon = 10^{-4}.$$

O erro relativo de $\bar{x}$ é pequeno. Será que o erro relativo em $f(\bar{x})$ também é pequeno? A resposta é negativa, porquanto

$$\delta_{\bar{x}} \simeq \frac{x - \bar{x}}{x} = \frac{-3\,\epsilon}{10 - \epsilon} \simeq -3 \times 10^{-1} \times \epsilon \ .$$

Todavia, dado que

$$f(x) = \frac{2}{x - 10} = -\frac{2}{\epsilon}$$

e

$$f(\bar{x}) = \frac{2}{\bar{x} - 10} = \frac{2}{2\,\epsilon} = \frac{1}{\epsilon},$$

obtém-se,

$$\delta_{f(\bar{x})} = \frac{f(x) - f(\bar{x})}{f(x)} = \frac{-3/\epsilon}{-2/\epsilon} = 3/2 = 1.5 \ .$$

Assim, o erro relativo do valor aproximado $f(\bar{x})$ é de cerca de 150 %, ou seja, esse valor estará completamente errado. ♦

Exemplo 1.9. *Considere-se agora a função de duas variáveis*

$$z = f(x, y) = x - y, \quad \textit{com} \quad x, y > 0 \ .$$

Esta função é bem condicionada para todos os pontos do seu domínio?

Da fórmula de propagação de erro (1.16), pág. 17, resulta

$$e_{\tilde{x}-\tilde{y}} = e_{\tilde{x}} - e_{\tilde{y}},$$

e aplicando fórmula de propagação do erro relativo (1.20), obtém-se

$$\delta_{\bar{x}-\bar{y}} \simeq \frac{x}{x-y}\,\delta_{\bar{x}} - \frac{y}{x-y}\,\delta_{\bar{y}}\ .$$

Analisemos, por exemplo, o número de condição relativamente à variável x,

$$cond_{f,1}(x,y)) = \frac{|x|}{|x-y|}\ .$$

Como

$$\lim_{x\to y} cond_{f,1}(x,y) = +\infty,$$

concluímos que a função f é mal condicionada para valores de x próximos de y (neste caso o número de condição $cond_{f,2}(x,y)$ é também ilimitado quando fazemos tender uma das variáveis para a outra).

O mau condicionamento desta função está na origem do fenómeno de *cancelamento subtractivo* a que se fez referência no Exercício 1.5, pág. 13.

♦

Convida-se o leitor a verificar que para $x, y > 0$, as seguintes funções z, de duas variáveis, $z = x + y$, $z = x \times y$ e $z = x/y$ são bem condicionadas, comparando as suas conclusões com a análise de propagação de erro efectuada na pág. 11.

Será interessante traçar os gráficos da função de condição de

$$\begin{array}{ll} f(x) = x^k, & \text{para } k \in \mathbb{N} \\ f(x) = \sqrt{x}, & x \geq 0 \\ f(x) = x^\alpha, & x > 0, \qquad 0 < \alpha < 1 \\ f(x) = \sin(x) & \\ f(x) = e^x. & \end{array}$$

Em particular, sabe-se que as funções trigonométricas de variável real são mal condicionadas para múltiplos de $\pi/2$. Tal circunstância obriga a que no desenvolvimento de software para essas funções se efectuem mudanças de base de representação numérica e redução a intervalos apropriados (sobre este problema ver, por exemplo, J. Harrison [15]).

Note-se finalmente que se nas fórmulas (1.19), (1.20), pág. 19, considerarmos os respectivos módulos, a majoração de erro assim obtida traduz a situação do chamado "pior caso", em que os erros envolvidos são todos do mesmo sinal, e portanto se adicionam uns aos outros.

Embora na prática computacional haja na realidade compensação de erros (os erros positivos compensando os negativos), deverá fazer-se uma análise considerando o *pior caso*, a fim de termos segurança absoluta quanto à precisão do resultado de um determinado cálculo, uma vez que a análise de erro levando em consideração essas compensações é geralmente difícil.

1.3 Propagação de erro em algoritmo

No parágrafo 1.1.5, pág. 11, ao referirmos a propagação de erros de arredondamento nas operações aritméticas elementares, admitimos que cada operação é efectuada exactamente, no sentido dado a seguir. Por exemplo, efectuemos o produto de dois números x e y, encarando o resultado como aplicação da função

$$z = f(x, y) = x \times y \ .$$

Se em vez dos valores exactos x e y, considerarmos valores $\bar{x}$ e $\bar{y}$, obtidos por arredondamento num sistema FP, sabemos que por aplicação da fórmula de propagação do erro relativo (1.19), ou (1.20), pág. 19, resulta a seguinte aproximação do erro propagado pela função,

$$\delta_{f(\bar{x},\bar{y})} \simeq \delta_{\bar{x}} + \delta_{\bar{y}} \ .$$

Acontece todavia que o resultado apresentado por um sistema $FP(\beta, n, t_1, t_2)$, não é em geral exactamente $\bar{z} = f(\bar{x}, \bar{y})$, mas antes $\bar{\bar{z}} = \bar{f}(\bar{x}, \bar{y})$, visto que

$$\bar{\bar{z}} = fl(fl(x) \times fl(y)),$$

(estamos assumindo que a operação $\times$ no sistema é efectuada exactamente). Há, portanto, que levar em consideração que o valor de $fl(x) \times fl(y)$ é geralmente arredondado antes de ser apresentado o resultado final $\bar{\bar{z}}$.

Faz por conseguinte sentido adoptar como modelo de propagação do erro relativo em cada passo de um algoritmo (subentendendo que nesse passo está em jogo uma certa função elementar $f(x)$, onde x é uma variável com um ou mais argumentos),

$$\delta_{\bar{f}(\bar{x})} = \frac{f(x) - \bar{f}(\bar{x})}{f(x)} \simeq \delta_{f(\bar{x})} + \delta_{arr}, \quad \text{com} \quad |\delta_{arr}| \leq u \ . \tag{1.21}$$

A primeira parcela no membro direito de (1.21) representa o erro relativo propagado pela função f (quando o argumento x é substituido por $\bar{x}$), enquanto que a parcela δ_{arr} representa o erro de arredondamento devido à operação em causa.

Ao efectuarmos um algoritmo de k passos, é eventualmente introduzido um erro relativo de arredondamento em cada passo, seja δ_{arr_i}, para $i = 1 : k$. O erro relativo do resultado em cada operação elementar pode ser muito ampliado em passos subsequentes. Neste caso dizemos que o algoritmo é numericamente *instável* para o conjunto de dados que servem de *input* ao algoritmo.

Relembre-se de que no Exemplo 1.5, pág. 13, foi usada uma função de uma variável, a qual exibia um comportamento instável para um certo valor do seu argumento. Esse mesmo exemplo é retomado a seguir.

Exemplo 1.10. *Considere de novo a função* $f(x) = (1 - \cos(x))/x^2$.

Reutilize o algoritmo descrito em (1.10)*, pág. 15, tendo por objectivo o cálculo de* $f(10^{-6})$*. Usando uma fórmula diferencial adequada, capaz de modelar o respectivo erro relativo propagado, estude a estabilidade numérica desse algoritmo.*
A função f *é bem condicionada para valores de* x *próximos de zero?*

Apliquemos o modelo de propagação de erro ao algoritmo de três passos a seguir.

$$\begin{array}{lll} z_1 = \cos(x) & \delta_{\bar{z}_1(\bar{x})} \simeq -\dfrac{x\ \sin(x)}{z_1}\,\delta_{\bar{x}} + \delta_{arr_1} \\ z_2 = 1 - z_1 & \delta_{\bar{z}_2(\bar{z}_1)} \simeq -\dfrac{z_1}{z_2}\,\delta_{\bar{z}_1} + \delta_{arr_2} \\ z_3 = \dfrac{z_2}{x^2} & \delta_{\bar{z}_3(\bar{x},\bar{z}_2)} \simeq\ \delta_{\bar{z}_2} - \delta_{\bar{x}^2} + \delta_{arr_3}\ . \end{array}$$

Substituindo sucessivamente as estimativas de erro obtidas em cada passo, obtém-se

$$\delta_{\bar{z}_2(\bar{z}_1)} \simeq -\frac{x\ \sin(x)}{1-\cos(x)}\,\delta_{\bar{x}} - \frac{\cos(x)}{1-\cos(x)}\delta_{arr_1} + \delta_{arr_2},$$

$$\delta_{\bar{z}_3(\bar{x},\bar{z}_2)} \simeq \left(\frac{x\ \sin(x)}{1-\cos(x)}\ - 2\right)\,\delta_{\bar{x}} - \frac{\cos(x)}{1-\cos(x)}\delta_{arr_1} + \delta_{arr_2} + \delta_{arr_3}\ .$$

Assim, uma majoração do erro relativo propagado ao algoritmo é,

$$\begin{array}{rl} \delta_{\bar{f}(\bar{x})} \leq & \left(|\dfrac{x\ \sin(x)}{1-\cos(x)}\ - 2|\right)\ |\delta_{\bar{x}}| + \\ & + \left(\dfrac{|\cos(x)|}{|1-\cos(x)|}\,|\delta_{arr_1}| + |\delta_{arr_2}| + |\delta_{arr_3}|\right). \end{array} \tag{1.22}$$

A primeira parcela do membro direito da desigualdade (1.22) reflecte o erro propagado pela função f (independentemente do algoritmo utilizado), enquanto a segunda parcela diz respeito ao erro de arredondamento propagado pelas sucessivas operações elementares que constituem o algoritmo.

No presente caso a função f é muito "bem comportada", porquanto o seu número de condição é

$$cond_f(x) = \left|\frac{x\ \sin(x)}{1-\cos(x)} - 2\right|.$$

Com efeito, atendendo a que

$$\begin{array}{rl} \lim_{x\to 0}\ cond_f(x) & = \left|-2 + \lim_{x\to 0}\,\dfrac{\sin(x) + x\ \cos(x)}{\sin(x)}\right| \\ & = -\left|1 + \lim_{x\to 0}\,\dfrac{x\ \cos(x)}{\sin(x)}\right| = 0, \end{array}$$

conclui-se que a função $f(x)$ é muito bem condicionada para valores de x próximos de zero, podendo mesmo contrair erros de arredondamento eventualmente cometidos, quando o seu argumento está próximo de zero. No entanto, atendendo à expressão (1.22), existe um *peso* afectando $|\delta_{arr_1}|$, tal que

$$\lim_{x\to 0} \frac{|\cos(x)|}{|1-\cos(x)|} = +\infty \ .$$

Assim, para valores de x próximos de zero, um pequeno erro relativo $|\delta_{\bar{x}}|$ no primeiro passo do algoritmo é muito ampliado no passo a seguir.

Note que no segundo passo, o cálculo de $z_2 = 1 - z_1$ corresponde a uma subtracção de números próximos, ou seja, ocorre o fenómeno de *cancelamento subtractivo* a que fizemos já referência (ver pág. 12).

Conclui-se assim que, para valores $x \simeq 0$, podemos em (1.22) negligenciar a parcela referente ao erro propagado pela função, mas não o podemos fazer quanto à parcela do erro devido ao algoritmo, obtendo-se

$$|\delta_{\bar{f}(\bar{x})}| \leq \frac{|\cos(x)|}{|1-\cos(x)|}\, u + 2\, u,$$

onde u é a unidade de arredondamento do sistema FP usado.

Admitindo, por exemplo, que o sistema decimal de representação numérica possui 10 dígitos na mantissa, a sua unidade de arredondamento simétrico é $u = 0.5 \times 10^{-9}$. Sendo $x = 10^{-6}$, se utilizarmos a fórmula anterior de majoração de erro propagado pelo algoritmo anteriormente considerado, resulta

$$|\delta_{\bar{f}(\bar{x})}| \leq 0.5 \times 10^{3},$$

ou seja, o erro relativo no resultado final será da ordem de $50\,000\,\%$, o que quer dizer que o resultado estará, como já tivemos oportunidade de constatar, completamente errado.

Uma vez que a função f é bem condicionada, para se calcular por exemplo $f(10^{-6})$, é forçoso substituir o algoritmo anterior por outro numericamente estável, tal como se fez no Exemplo 1.5, pág. 13.

O Exemplo (1.10) mostra-nos que para resolver um problema concreto é desejável dispor de vários algoritmos distintos, porquanto algum deles pode ser numericamente instável para o conjunto de dados usados no problema em causa.

Ao longo do curso teremos oportunidade de tomar contacto com algoritmos que à primeira vista são muito apelativos para resolver um determinado problema, mas que não serão utilizados na prática devido à sua instabilidade numérica.
Por razões óbvias, se uma dada função f for mal condicionada para um certo conjunto de dados, todo e qualquer algoritmo construído para a calcular estará sujeito a instabilidade numérica. Nesse caso, ou se reformula completamente o problema, ou seremos forçados a usar cálculos com precisão aumentada.

♦

1.4 Leituras aconselhadas

David Goldberg, *What Every Computer Scientist Should Know About Floating-Point Arithmetic*, Computing Surveys, ACM, 1991.

(Disponível em Institute of Electrical and Electronics Engineers, New York, http://grouper.ieee.org/groups/754).

John Harrison, *Decimal transcendentals via binary*, Computer Arithmetic, IEEE, 187-194, 2009.

Capítulo 2

Métodos numéricos para equações não lineares

2.1 Raízes de equações não lineares

Equações não lineares, do tipo $f(x) = 0$ ou $x = h(x)$, surgem naturalmente nas aplicações quando um determinado fenómeno físico é modelado matematicamente usando um determinado princípio de equílibrio. Por exemplo, sob certas condições, pode deduzir-se da segunda lei de Newton[1], que a velocidade $v(x)$ de um corpo em queda livre satisfaz a seguinte equação não linear, na variável x,

$$v(x) = \frac{m}{\alpha}\, g\, \left(1 - e^{-\frac{\alpha}{m}\, x}\right),$$

onde α, m e g são constantes ou parâmetros dependentes do sistema físico em causa.

Pode colocar-se a questão de saber como determinar o parâmetro α na equação anterior (α representa um coeficiente de resistência do ar), caso se conheçam os valores de x, $v(x)$ e dos restantes parâmetros. Podemos reescrever essa equação, por exemplo, na forma

$$\alpha = \frac{m}{v(x)}\, g\, \left(1 - e^{-\frac{\alpha}{m}\, x}\right), \tag{2.1}$$

ou

$$\alpha - \frac{m}{v(x)}\, g\, \left(1 - e^{-\frac{\alpha}{m}\, x}\right) = 0\ .$$

Assim, determinar um valor α satisfazendo a equação (2.1) equivale a "resolver"uma das equações equivalentes

$$\alpha = h(\alpha), \quad \text{com} \quad h(\alpha) = \frac{m}{v(x)}\, g\, \left(1 - e^{-\frac{\alpha}{m}\, x}\right),$$

[1] Isaac Newton, 1642-1727, físico e matemático inglês, considerado um dos maiores cientistas de todos os tempos.

ou

$$f(\alpha) = 0, \quad \text{com} \quad f(\alpha) = \alpha - \frac{m}{v(x)}\, g\, \left(1 - e^{-\frac{\alpha}{m}\, x}\right).$$

Neste capítulo discutiremos como "resolver" uma equação real não linear do tipo anteriormente considerado, ou seja, da forma $f(x) = 0$ ou $x = g(x)$, onde f e g são funções dadas de variável real.

No conjunto das equações não lineares numa variável real x, avultam as equações polinomiais. Um polinómio oferece a vantagem de ser facilmente calculável num ponto, ser uma função regular (no sentido em que existem e são contínuas as suas derivadas de qualquer ordem), as suas derivadas são facilmente calculáveis, e o integral de um polinómio pode igualmente ser facilmente obtido. Todavia, determinar o conjunto solução para uma equação polinomial $f(x) = 0$, pode não ser tarefa fácil.

Uma exposição detalhada e interessante a respeito da evolução dos algoritmos para o cálculo numérico de raízes de equações encontra-se em Knoebel et al. [23].

Comecemos por definir o que se entende por *zero* de uma função. Seja f uma função real, definida num certo intervalo $[a, b]$. O ponto $z \in [a, b]$ diz-se um *zero* de f, ou uma *raiz* da equação $f(x) = 0$ se $f(z) = 0$.

Admitindo que uma função f é suficientemente regular, classificamos um seu zero como *simples* ou *múltiplo*, de acordo com a definição a seguir.

Definição 2.1. Sendo $f(z) = 0$ e $f'(z) \neq 0$, o zero z diz-se *simples*. Se $f'(z) = 0$, z diz-se um zero *múltiplo*. Mais precisamente, se $f \in C^k(z)$ e se

$$f'(z) = f''(z) = ... = f^{(k-1)}(z) = 0 \quad \text{e} \quad f^{(k)}(z) \neq 0,$$

z diz-se um zero de multiplicidade k da função f.

Exemplo 2.1. *Seja f um polinómio de grau n, com $n \geq 1$. De acordo com o teorema fundamental da álgebra, o polinómio possui n raízes em $\mathbb{C}$ (somando as suas multiplicidades).*

(a) A função polinomial $f(x) = x^k$, $k \geq 1$, possui um só zero real, $z = 0$, de multiplicidade k. Todas as derivadas de f são nulas em $z = 0$, excepto a de ordem k, para a qual $f^{(k)}(x) = k!$.

(b) Se tivermos, por exemplo, um polinómio do segundo grau,

$$f(x) = x^2 + 2\,x + 1 = (x + 1)^2,$$

este polinómio possui uma raiz de multiplicidade dois (raiz dupla) em $z = -1$. De facto, verifica-se a igualdade $f(-1) = 0$. Visto que $f'(x) = 2\,x + 2$, temos $f'(-1) = 0$. Como $f''(x) = 2$, resulta $f''(-1) \neq 0$.

(c) Se considerarmos a equação polinomial de terceiro grau

$$f(x) = x^3 - x = x\,(x-1)\,(x+1) = 0,$$

existem três raízes simples: $z_1 = -1$, $z_2 = 0$ *e* $z_3 = 1$.

(d) O polinómio

$$f(x) = x^3 + 1,$$

possui apenas uma raiz real ($z_1 = -1$) *e duas raízes complexas conjugadas* ($z_{2,3} = \dfrac{1 \pm \sqrt{3}\,i}{2}$).

♦

De um modo geral, a determinação dos zeros de um polinómio de grau $n \geq 1$, de coeficientes reais (ou seja, as raízes de uma equação algébrica), é um problema complexo que ocupou os matemáticos de várias épocas.

Desde o início do século XIX sabe-se, graças a Abel[2], que não existem fórmulas resolventes para equações algébricas em geral. Mais precisamente, para uma equação algébrica de grau superior a 4, não é possível exprimir as suas raízes através dos coeficientes do polinómio mediante fórmulas envolvendo somas, subtracções, multiplicações, divisões e radicais.

Tal circunstância ilustra a importância dos métodos numéricos para a resolução de equações. Até no caso de equações relativamente simples, como as equações algébricas, é geralmente impossível calcular as suas raízes através de fórmulas analíticas. Por outro lado, mesmo nos casos em que existem fórmulas resolventes, estas são por vezes tão complexas que se torna mais eficiente determinar as raízes a partir de um método numérico. Tal é o caso de algumas equações algébricas de terceiro e quarto graus, por exemplo. Naturalmente, isso pressupõe que se escolha um método numérico adequado.

A evolução do pensamento matemático no tratamento de equações algébricas é magistralmente tratada por John Stillwell em [34]. Retenha-se o que nos ensina este autor: "The most fertile problems in mathematics are over 2000 years old and still have not yielded up all their secrets "([34], p. 1).

Equações não algébricas dir-se-ão *transcendentes*. O exemplo a seguir leva-nos a tentar "resolver" uma certa equação transcendente.

Exemplo 2.2. *A Figura 2.1 representa o perfil de um determinado terreno onde se encontra instalado um cabo eléctrico ligando dois pontos A e B.*

Pretende-se determinar a altura h que medeia entre o ponto C e o ponto mais baixo do cabo figurado. Conhecem-se as distâncias d, L e b.

[2]Niels Henrik Abel, 1802-1829, matemático norueguês.

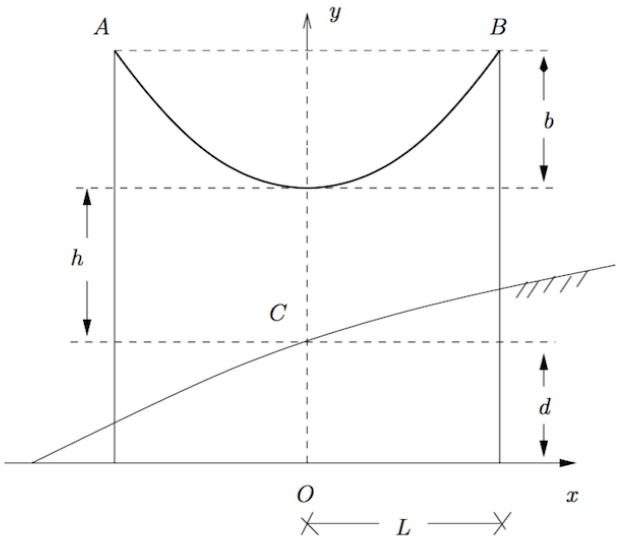


Figura 2.1: O problema da catenária ligando os pontos A e B.

Sabe-se que (relativamente ao referencial xOy na figura) a linha que representa o cabo tem como expressão analítica

$$y(x) = a \cosh\left(\frac{x}{5\,a}\right),$$

onde $a = d + h$, e cosh *designa a função coseno hiperbólico, definida em $\mathbb{R}$ pela expressão* $\cosh(t) = (e^t + e^{-t})/2$.

A linha considerada é conhecida pela designação de catenária, e o parâmetro a diz-se o parâmetro da catenária.

Supondo que $d = 10\,m$, $L = 100\,m$ e $b = 5\,m$, qual é o parâmetro da catenária em causa, e a respectiva altura h?

Atendendo a que $y(L) = h + d + b = a + b$, tem-se

$$a \cosh\left(\frac{L}{5\,a}\right) = a + b\ .$$

A equação anterior é equivalente a

$$a \cosh\left(\frac{L}{5\,a}\right) - a - b = 0\ .$$

Por conseguinte, o parâmetro a da catenária será um zero da função,

$$f(a) = a \cosh\left(\frac{L}{5\,a}\right) - a - b\ .$$

A altura $h = a - d$, será um zero da função

$$\begin{aligned}\phi(h) &= (d+h)\,\cosh\left(\frac{L}{5\,(d+h)}\right) - (d+h) - b \\ &= (60+h)\,\cosh\left(\frac{20}{60+h}\right) - 65 + h\ .\end{aligned}$$

♦

O problema proposto no Exemplo 2.2 sugere que existe raiz real positiva para a equação $f(a) = 0$, ou equivalentemente para $\phi(h) = 0$, e que tal raiz é única. Surgem então naturalmente as seguintes questões:

- Provar que as referidas equações possuem solução z e que a solução é única;
- Localizar z;
- Calcular z com erro absoluto, por exemplo, não superior a $1\,cm$.

Nos parágrafos seguintes discutiremos a teoria que nos habilita a responder a questões análogas, quanto se pretende resolver uma qualquer equação não linear do tipo $f(x) = 0$ ou $x = g(x)$, onde f e g são funções dadas. No Capítulo 3 lidaremos com o problema mais complexo respeitando ao cálculo de aproximações de raízes de *sistemas de equações* não lineares (ver pág. 168).

2.1.1 Localização de raízes

Para tratar o problema do cálculo numérico das raízes de uma dada equação $f(x) = 0$, é necessário em primeiro lugar localizá-las, isto é, determinar para cada raiz um intervalo que a contenha e não contenha nenhuma outra.

Com esse objectivo, recordemos dois teoremas da análise matemática associados respectivamente a B. Bolzano[3] e M. Rolle[4] (para a sua demonstração ver, por exemplo [14]).

Teorema 2.1. (Teorema de Bolzano)
Se f for contínua em $[a, b]$ e se $f(a)f(b) < 0$, então f possui *pelo menos* uma raiz em (a, b).

Teorema 2.2. (Corolário do teorema de Rolle)
Se f for contínua em $[a, b]$, continuamente diferenciável em (a, b), e se $f'(x) \neq 0$ em (a, b), então f possui *no máximo* uma raiz em(a, b) .

[3]Bernhard Bolzano, 1781-1848, matemático e teólogo, natural da Boémia.
[4]Michel Rolle, 1652-1719, matemático francês.

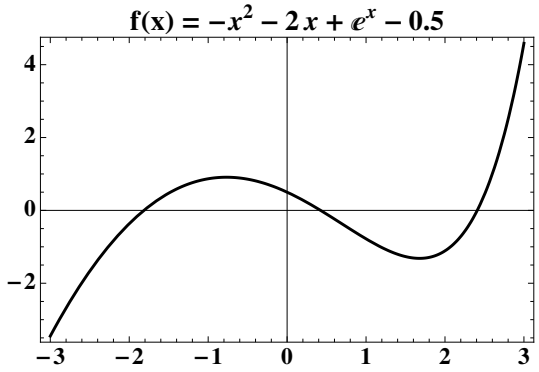


Figura 2.2: Gráfico relativo ao Exemplo 2.3 (três raízes reais simples).

Combinando estes dois teoremas com outros resultados da Análise, é possível, em muitas situações, localizar as raízes reais de uma equação.

Um outro teorema fundamental que teremos oportunidade de usar com frequência, é o chamado Teorema de Lagrange[5], que aqui se relembra.

Teorema 2.3. (Teorema de Lagrange)
Se f é uma função contínua no intervalo $[a, b]$ e diferenciável em (a, b), existe pelo menos um ponto $\xi \in (a, b)$, tal que

$$f(b) = f(a) + f'(\xi)\,(b - a).$$

Demonstração. Ver, por exemplo, [14], pág. 380. □

Exemplo 2.3. *Com base nos Teoremas 2.1 e 2.2, determinar o número de raízes reais da equação*

$$e^x - x^2 - 2\,x = 0.5,$$

e obter para cada uma delas um intervalo que apenas contenha essa raiz.

Este problema é equivalente a determinar os zeros da função de variável real $f(x) = e^x - x^2 - 2\,x - 0.5$. A função é evidentemente contínua em $\mathbb{R}$, assim como todas as suas derivadas de qualquer ordem. Pode observar na Figura 2.2 que os zeros de f (pontos de intersecção do gráfico da função com o eixo x das abcissas), grosso modo estão próximos de -2, 0 e 3. Como nesses zeros a função derivada f' é não nula, os zeros são simples (ver Definição 2.1, pág. 28).

Para facilitar a análise do problema, comecemos por calcular os seguintes valores de f e de f', indicados na tabela seguinte.

x	-3	-2	-1	0	1	2	3
$f(x)$	-3.45	-0.365	0.868	0.5	-0.782	-1.11	4.59
$f'(x)$	4.05	2.14	0.368	-1	-1.28	1.39	12.1

[5]Joseph-Louis Lagrange, 1736 -1813, matemático e astrónomo, nascido em Itália.

Observando a tabela anterior verifica-se imediatamente que o Teorema 2.1 é aplicável à função f nos intervalos $[-2,-1]$, $[0,1]$ e $[2,3]$. Daqui se conclui que a equação considerada possui pelo menos três raízes reais, respectivamente $z_1 \in [-2,-1], z_2 \in [0,1]$ e $z_3 \in [2,3]$.

Pelo Teorema 2.2 podemos concluir também que, em cada um desses intervalos, a função f possui exactamente uma raiz. De facto, consideremos as derivadas

$$f'(x) = e^x - 2x - 2, \qquad f''(x) = e^x - 2 \ .$$

Em relação à segunda derivada, verifica-se facilmente que ela é positiva para $x > \ln 2$ e negativa para $x < \ln 2$.
Temos $f''(\ln 2) = 0$ e $f'''(\ln 2) = 2$, pelo que f' tem em $x = \ln 2$ um ponto de mínimo. Assim, no intervalo $[-2,-1]$ a função f' é decrescente.
Recorrendo de novo à tabela verifica-se que f' é sempre positiva neste intervalo. Pelo Teorema 2.2, podemos concluir que f possui um único sero z_1 no intervalo $[-2,-1]$.

Do mesmo modo podemos observar que a função f' é crescente em $[2,3]$ e, de acordo com a tabela, toma sempre valores positivos neste intervalo. Aplicando o Teorema 2.2 neste intervalo, constata-se que f tem nele um único zero, seja z_3.

Para se aplicar o mesmo teorema no intervalo $[0,1]$, comecemos por recordar que a função f' tem um ponto de mínimo em $x = \ln 2$, que pertence a este intervalo. Note-se que $f'(\ln 2) = -1.38 < 0$, e de acordo com a tabela anterior, $f'(0)$ e $f'(1)$ também são negativos, pelo que podemos concluir ser f' negativa em todo o intervalo $[0,1]$. Logo, o Teorema 2.2 é aplicável neste intervalo e a função tem nele um único zero z_2.

Resta esclarecer uma questão: será que a equação $f(x) = 0$ possui alguma raiz real além das que acabámos de localizar? Para responder a esta pergunta, recordemos que a segunda derivada de f tem uma única raiz real em $x = \ln 2$. Pelo Teorema de Rolle somos levados a concluir que a primeira derivada de f tem, no máximo, duas raízes reais. Finalmente, aplicando o Teorema de Rolle a f', conclui-se que f possui no máximo três raízes reais. Como já vimos que existem pelo menos três raízes (z_1, z_2 e z_3), concluímos que estas são as únicas raízes da equação $f(x) = 0$. ⧫

2.1.2 Estimativas de erro

Considere-se uma função f continuamente diferenciável num intervalo $I = [\alpha, \beta]$, e z um zero de f isolado em I, tal que

$$f'(x) \neq 0, \qquad \forall x \in I \ . \tag{2.2}$$

Designando por x_k uma dada aproximação de z, por $int(x_k, z)$ o intervalo aberto $(min\{x_k, z\}, max\{x_k, z\})$ e por e_k o erro da aproximação, $e_k = z - x_k$, segundo o

Teorema 2.3, com $a = x_k$ e $b = z$, podemos garantir a existência de (pelo menos) um ponto ξ_k tal que

$$0 = f(z) = f(x_k) + f'(\xi_k)\,(z - x_k), \qquad \xi_k \in int(x_k, z)\ . \tag{2.3}$$

Uma vez que por hipótese a função f' é não nula em qualquer ponto do intervalo I, é positivo o seguinte valor de mínimo,

$$m = \min_{\alpha \le x \le \beta} |f'(x)| > 0\ . \tag{2.4}$$

Além disso, supondo x_k suficientemente próximo de z, dado que $\xi_k \in int(x_k, z)$ e f' contínua, tem-se que

$$f'(x_k) \simeq f'(\xi_k) \neq 0\ .$$

Por conseguinte, atendendo a (2.3), obtém-se a estimativa de erro $\bar{e}_k \simeq z - x_k$, da forma

$$\bar{e}_k = -\frac{f(x_k)}{f'(x_k)}\ . \tag{2.5}$$

De (2.3), levando em consideração a desigualdade em (2.4), resulta a majoração do erro de x_k,

$$|z - x_k| = |e_k| \le \frac{|f(x_k)|}{min_{\alpha \le x \le \beta}\,|f'(x)|} = \frac{|f(x_k)|}{m}\ . \tag{2.6}$$

Note-se que para aplicarmos a majoração de erro anterior é imprescindível admitirmos a hipótese (2.2). Com efeito, por exemplo para a função $f(x) = x^2$ e um qualquer intervalo $I = [-\alpha, \alpha]$, com $\alpha > 0$, existe um só zero $z = 0$ no intervalo, e $m = min_{-\alpha \le x \le \alpha}|f'(x)| = f'(z) = 0$. No entanto, sendo $x_k \neq 0$ uma aproximação de $z = 0$ nesse intervalo, a estimativa de erro (2.5) é válida, resultando

$$\bar{e}_k = -\frac{x_k^2}{2\,x_k} = \frac{x_k}{2}\ .$$

Exemplo 2.4.

Nas condições referidas anteriormente, a estimativa de erro (2.5) sugere-nos um método clássico para obter aproximações cada vez melhores de um zero simples z de uma função f. Tal método será estudado na Secção 2.3, página 65.

Relembre-se (ver Exemplo 2.3) que a função

$$f(x) = e^x - x^2 - 2\,x - 1/2,$$

tem um único zero $z \in [-2,-1]$. Sendo x_k uma aproximação de z, atendendo a que $z - x_k \simeq \bar{e}_k$, onde $\bar{e}_k$ é a estimativa de erro dada em (2.5), é natural designar-se por x_{k+1} a seguinte nova aproximação de z,

$$x_{k+1} = x_k + \bar{e}_k = x_k - \frac{f(x_k)}{f'(x_k)} \ . \tag{2.7}$$

Por exemplo, xeja $x_0 = -2$ uma primeira aproximação de z . Algumas aplicações sucessivas de (2.5), (2.6) e (2.7) são indicadas a seguir. Para a função em causa tem-se

$$m = \min_{-2\leq x\leq -1} |f'(x)| = f'(-1) = 1/e \ .$$

Os valores das aproximações x_1, x_2 e x_3 abaixo foram obtidos mediante arredondamento simétrico para 10 casas decimais.

$$x_0 = -2$$

$$\bar{e}_0 = -\frac{f(-2)}{f'(-2)} = 0.1707763270\cdots \quad \text{(estimativa de erro)}$$

$$|z - x_0| \leq \frac{|f(x_0)|}{m} = 0.991\cdots \quad \text{(majoração de erro)} \ .$$

Para

$$x_1 = x_0 + \bar{e}_0 = -1.8292236729, \quad \text{(nova aproximação)}$$

$$\bar{e}_1 = -\frac{f(x_1)}{f'(x_1)} = 0.0148839835\cdots$$

$$|z - x_1| \leq \frac{|f(x_1)|}{m} = 0.0735\cdots \ .$$

Para

$$x_2 = x_1 + \bar{e}_1 = -1.8143396894,$$

$$\bar{e}_2 = -\frac{f(x_2)}{f'(x_2)} = 0.0001137645\cdots$$

$$|z - x_2| \leq \frac{|f(x_2)|}{m} = 0.000553\cdots \ .$$

Para

$$x_3 = x_2 + \bar{e}_2 = -1.8142260148,$$

$$\bar{e}_3 = -\frac{f(x_3)}{f'(x_3)} = 6.62\cdots 10^{-9}$$

$$|z - x_3| \leq \frac{|f(x_3)|}{m} = 3.22\cdots 10^{-8} \ .$$

Conclui-se que o número $\bar{z} = -1.8142260$ é uma aproximação do zero z, onde todos os seus dígitos são significativos.

♦

2.1.3 Método da bissecção

Um dos métodos mais simples para o cálculo aproximado de raízes é o método da bissecção. Para se poder aplicar este método basta que a função em causa seja *contínua* num intervalo que contenha uma única raiz da função.

A ideia do método é construir uma sucessão de intervalos encaixados,

$$[a, b] \supset [a_1, b_1] \supset ... \supset [a_k, b_k],$$

tais que:

a) Cada intervalo tem o comprimento igual a metade do intervalo anterior;

b) Em cada intervalo é satisfeita a condição $f(a_i)f(b_i) < 0, \qquad i = 1 : k.$

O Teorema 2.1, pág. 31, sugere que a raiz é um ponto comum a todos os intervalos da sucessão. Assim, se considerarmos um número suficientemente grande de intervalos, é possível aproximar a raiz com a precisão que se pretender.

Vejamos em pormenor o algoritmo deste método.

1º Passo

Dado um intervalo $[a, b]$, e uma função f tais que $f(a)f(b) < 0$, determina-se o ponto médio desse intervalo $x_1 = \dfrac{a+b}{2}$.
Se, por coincidência, se verificar $f(x_1) = 0$, o ponto x_1 é a raiz procurada e o processo termina. Suponhamos que $f(x_1) \neq 0$. Então, verifica-se

$$f(x_1)f(a) < 0 \quad \text{ou} \quad f(x_1)f(a) > 0.$$

No primeiro caso, podemos afirmar que a raiz $z \in [a, x_1]$, no segundo caso $z \in [x_1, b]$. Assim, o intervalo $[a_1, b_1]$ pode ser definido do seguinte modo:

Se $f(x_1)f(a) < 0$, então fazer $a_1 = a$ e $b_1 = x_1$; caso contrário, fazer $a_1 = x_1$ e $b_1 = b$.
Em qualquer dos casos, o novo intervalo $[a_1, b_1]$ satisfaz $f(a_1)f(b_1) < 0$.

2º Passo

Repetem-se as acções do primeiro passo, substituindo o intervalo $[a, b]$ por $[a_1, b_1]$, e representando por x_2 o ponto médio deste intervalo. O resultado deste passo é o intervalo $[a_2, b_2]$.

Generalizando, no *k-ésimo* passo (iteração), procede-se do seguinte modo:

Determina-se o ponto médio do intervalo anterior,

$$x_k = \frac{a_{k-1} + b_{k-1}}{2}. \tag{2.8}$$

Se $f(x_k)f(a_{k-1}) < 0$, então fazer $a_k = a_{k-1}$ e $b_k = x_k$; senão fazer $a_k = x_k$ e $b_k = b_{k-1}$. No k-*ésimo* passo obtém-se o intervalo $[a_k, b_k]$.

O processo é interrompido quando for satisfeita a *condição de paragem*

$$b_k - a_k < \varepsilon,$$

onde ε é uma tolerância previamente estabelecida, de acordo com a precisão que se pretende obter.

Estimativas de erro

Note-se que o comprimento do k-ésimo intervalo, por construção, vale

$$b_k - a_k = \frac{b-a}{2^k},$$

pelo que esse valor tende para zero, quando k tende para infinito. Logo, qualquer que seja a tolerância ε, a condição de paragem é satisfeita ao fim de um certo número de passos (dependendo do comprimento do intervalo inicial e de ε). Mais precisamente, temos

$$\frac{b-a}{2^k} < \varepsilon \iff \frac{b-a}{\varepsilon} < 2^k \iff k > \log_2\left(\frac{b-a}{\varepsilon}\right).$$

Assim, o número de passos do método da bissecção que é necessário realizar até satisfazer a condição de paragem é o menor inteiro k, tal que

$$k > \log_2\left(\frac{b-a}{\varepsilon}\right).$$

Se tomarmos como k-ésima aproximação da raiz z o valor de x_k, podemos afirmar que o erro absoluto de x_k satisfaz a desigualdade

$$|z - x_k| < \frac{b_{k-1} - a_{k-1}}{2} = \frac{b-a}{2^k}.$$

Nada impede que denotemos por x_0 o extremo a ou b do intervalo inicial. Nesse caso, por construção do método, é válida a relação

$$\frac{b-a}{2^k} = |x_k - x_{k-1}|.$$

É costume nos métodos computacionais representar o erro da k-ésima aproximação da raiz por e_k. Usando esta notação, podemos afirmar que no método da bissecção são válidas as majorações de erro

$$\begin{array}{ll} |e_k| = |z - x_k| & < \dfrac{b-a}{2^k}, \\ \text{ou} & \\ |e_k| = |z - x_k| & < |x_k - x_{k-1}|, \qquad k = 1, 2, \ldots . \end{array} \tag{2.9}$$

Convergência

Mostremos que, de facto, o método converge para a solução de $f(x) = 0$.

Por construção do método, sabemos que

$$f(a_k) \times f(b_k) < 0, \qquad k = 1, 2, \ldots \tag{2.10}$$

e que

$$a_{k-1} < x_k < b_{k-1}, \qquad k = 1, 2, \ldots \tag{2.11}$$

A sucessão $(a_{k-1})_{k\geq 0}$ é monótona não decrescente limitada por $b_0 = b$, e a sucessão $(b_{k-1})_{k\geq 0}$ é monótona não crescente limitada por $a_0 = a$. Por conseguinte, estas sucessões são convergentes.

Sejam $\alpha = \lim_{k\to\infty}(a_{k-1})$ e $\beta = \lim_{k\to\infty}(b_{k-1})$. Atendendo à desigualdade (2.11), tem-se

$$\alpha \leq \lim_{k\to\infty} x_k \leq \beta.$$

Mas, como $b_k - a_k < \dfrac{b-a}{2^k}$, para $k = 0, 1, \ldots$, resulta $\lim_{k\to\infty}(b_k - a_k) = 0$ e

$$\lim_{k\to\infty}(b_k - a_k) = \beta - \alpha \iff \alpha = \beta \ .$$

Quer isto dizer que as sucessões constituídas respectivamente pelos extremos dos subintervalos $[a_k, b_k]$ são ambas convergentes para o mesmo número, e de (2.11) temos também

$$\lim_{k\to\infty} x_k = \alpha = \beta \ .$$

Seja z o limite comum anterior. Da desigualdade (2.10) e atendendo a que f é, por hipótese, *contínua*, obtém-se,

$$f\left(\lim_{k\to\infty} a_k\right)\ f\left(\lim_{k\to\infty} b_k\right) \leq 0,$$

isto é,

$$f^2(z) \leq 0 \ .$$

A desigualdade anterior é válida se e só se $f(z) = 0$. Como por hipótese só existe um zero de f em $[a, b]$, provámos que $\lim_{k\to\infty} x_k = z$.

A conclusão anterior de que $\lim_{k\to\infty} x_k = \alpha = \beta$, baseia-se no pressuposto de que a desigualdade (2.10) é válida para qualquer iterada k. No entanto, devido às limitações impostas pelo cálculo numérico, pode acontecer que para $k > k_0$, se verifique $\bar{f}(a_k) = 0$ e/ou $\bar{f}(b_k) = 0$, onde $\bar{f}$ representa o valor de f arredondado pelo sistema de ponto flutuante usado. Por conseguinte, deverá tomar-se a referida desigualdade como teórica, porquanto a sua validade fica limitada por eventuais erros de arredondamento cometidos pelo sistema de ponto flutuante utilizado. Mas, como em geral as aproximações a determinar para as iteradas x_k do método da bissecção estão ainda longe da solução exacta z, os respectivos valores calculados de $f(x_k)$ estarão por sua vez suficientemente longe de zero, pelo que uma avaliação incorrecta do sinal do produto $\bar{f}(a_k) \times \bar{f}(b_k)$ será uma situação excepcional.

Exemplo 2.5. *a) Recorrendo ao Teorema 2.1, pág. 31, justifique que a raiz cúbica de 2 pertence ao intervalo* $[1.2, 1.3]$.

b) Baseando-se na alínea anterior, efectue três iterações (passos) do método da bissecção, com o objectivo de calcular um valor aproximado de $\sqrt[3]{2}$.

c) Quantas iterações teria que efectuar se pretendesse determinar $\sqrt[3]{2}$ *com um erro absoluto inferior a* 0.001*?*

Comecemos por observar que determinar a raiz cúbica de 2 equivale a resolver a equação $f(x) = x^3 - 2 = 0$.

a) Temos que $f(1.2) = 1.2^3 - 2 = -0.272 < 0$ e $f(1.3) = 1.3^3 - 2 = 0.197 > 0$. Uma vez que a função f é contínua, pelo Teorema 2.1 concluímos que a raiz procurada está no intervalo $[1.2, 1.3]$.

b) Partindo do intervalo $[a, b] = [1.2, 1.3]$, a primeira iterada é $x_1 = \dfrac{a+b}{2} = 1.25$. Verifica-se que $f(1.25) = -0.047 < 0$, donde

$$f(1.25)f(1.2) > 0 \ .$$

Logo, o intervalo a considerar na iteração seguinte é $[a_1, b_1] = [1.25, 1.3]$. Por conseguinte, $x_2 = \dfrac{a_1 + b_1}{2} = 1.275$. Neste caso, $f(1.275) = 0.0727 > 0$, donde $f(1.275)f(1.25) < 0$. Assim, o intervalo a considerar na terceira iteração é $[a_2, b_2] = [1.25, 1.275]$. Finalmente, $x_3 = \dfrac{a_2 + b_2}{2} = 1.2625$.
Neste ponto, temos $f(1.2625) = 0.012 > 0$, pelo que o intervalo a considerar na iteração seguinte será $[a_3, b_3] = [1.25, 1.2625]$.

k	a_k	b_k	Sign(f(a_k))	Sign(b_k)	x_k	Sign(f(x_k))
0	20	50	1	-1	35 = 35.	1
1	35	50	1	-1	$\frac{85}{2}$ = 42.5	-1
2	35	$\frac{85}{2}$	1	-1	$\frac{155}{4}$ = 38.75	1
3	$\frac{155}{4}$	$\frac{85}{2}$	1	-1	$\frac{325}{8}$ = 40.625	1
4	$\frac{325}{8}$	$\frac{85}{2}$	1	-1	$\frac{665}{16}$ = 41.5625	-1
5	$\frac{325}{8}$	$\frac{665}{16}$	1	-1	$\frac{1315}{32}$ = 41.0938	-1
6	$\frac{325}{8}$	$\frac{1315}{32}$	1	-1	$\frac{2615}{64}$ = 40.8594	-1
7	$\frac{325}{8}$	$\frac{2615}{64}$	1	-1	$\frac{5215}{128}$ = 40.7422	1
8	$\frac{5215}{128}$	$\frac{2615}{64}$	1	-1	$\frac{10\,445}{256}$ = 40.8008	1
9	$\frac{10\,445}{256}$	$\frac{2615}{64}$	1	-1	$\frac{20\,905}{512}$ = 40.8301	-1
10	$\frac{10\,445}{256}$	$\frac{20\,905}{512}$	1	-1	$\frac{41\,795}{1024}$ = 40.8154	-1

Figura 2.3: Método da bissecção para o problema da catenária.

c) O comprimento do intervalo inicial é $b - a = 0.1$. Assim, para se atingir uma precisão de $\varepsilon = 0.001$, o número de iterações será

$$\log_2\left(\frac{b-a}{\varepsilon}\right) = \log_2\left(\frac{0.1}{0.001}\right) = 6.64.$$

Ou seja, a precisão pretendida será seguramente atingida ao fim de 7 iterações. ♦

O método da bissecção tem a vantagem de convergir, sempre que num intervalo $[a, b]$ se encontrar um zero isolado de uma função contínua nesse intervalo que mude de sinal nos extremos do intervalo. Porém, este método é geralmente de convergência lenta. Daí que ele seja frequentemente usado para obter uma estimativa "suficientemente próxima" de z. Tal estimativa é depois utilizada como *aproximação inicial* de z, tendo em vista a aplicação de um método numérico que convirja mais rapidamente do que o método da bissecção.

Exemplo 2.6. *Aplique o método da bissecção para obter uma estimativa inicial do parâmetro da catenária dada no Exemplo 2.2, pág. 29.*

Substituindo L e b pelos valores dados, a equação a resolver é

$$f(a) = a\,\cosh\left(\frac{20}{a}\right) - a - 5 = a\,\frac{e^{20/a} + e^{-20/a}}{2} - a - 5\ .$$

O problema pressupõe que $a > 0$. A função f é continuamente diferenciável.

Dado que $\lim_{a\to 0^+} f(a) = +\infty$, e

$$\lim_{a\to+\infty} f(a) = \lim_{a\to+\infty} \frac{\cosh(20\times a^{-1})-1}{\frac{1}{a}} - 5 = -5,$$

conclui-se que existe pelo menos uma raiz positiva da equação. Como

$$f'(a) = \cosh(20/a) - 20/a\,\sinh(20/a) - 1,$$

e

$$\begin{aligned} f''(a) &= -20/a^2\,\sinh(20/a) + 20^2/a^2\,\cosh(20/a) \\ &= \frac{400}{a^3}\,\cosh(20/a) > 0, \quad \forall a > 0, \end{aligned}$$

a função derivada f' é estritamente crescente e mantém sinal (negativo) em $\mathbb{R}^+$, logo f possui no máximo um zero real positivo. Atendendo a que

$$f(20) \simeq 5.9 > 0 \quad \text{e} \quad f(50) \simeq -0.95 < 0,$$

é certo que no intervalo $[20, 50]$ existirá o único zero positivo da função, prevendo-se que esse zero esteja mais próximo do valor 50 do que do valor 20.

Na Fig. 2.3 mostra-se o resultado da aplicação do método da bissecção no intervalo considerado. Pode observar-se a lentidão do processo – no final de 10 iterações o valor calculado $z \simeq 40.8154$, possui apenas 3 algarismos significativos. Na realidade z é aproximadamente 40.8071, como poderá concluir se usar um método de convergência rápida.

Chama-se a atenção de que a iterada x_8 (ver Fig. 2.3) é mais precisa do que x_{10}. Tal deve-se ao facto do método apenas analisar o *sinal*[6] da função em cada iterada, $sgn(x_k)$, comparando-o com o sinal da função num dos extremos do intervalo a partir do qual essa iterada é calculada.

Como veremos adiante, métodos usando mais informação sobre a função, quando convergentes, convergem em geral mais rapidamente do que o método aqui tratado.

Se, por exemplo, pretendêssemos aproximar a raiz z com uma tolerância $\epsilon < 10^{-6}$, o número de iterações a efectuar seria superior a 20. Com efeito, designando por N esse número, tem-se

$$|e_k| < \frac{b-a}{2^k} < \epsilon \iff 2^k > \frac{30}{\epsilon},$$

ou seja,

$$k > \frac{\log(30/\epsilon)}{\log(2)} \simeq 24.8\ .$$

[6] A função $sgn(x)$ define-se como $sgn(0) = 0$, $sgn(x) = 1$, se $x > 0$, e $sgn(x) = -1$, se $x < 0$.

Assim, se efectuarmos $N = 25$ iterações podemos garantir que o erro absoluto $|e_{25}| = |z - x_{25}| < 10^{-6}$. Este número de iterações pode ser considerado insignificante apenas se estivermos lidando com um cálculo isolado de uma raiz.

Nas aplicações são frequentes os modelos matemáticos para os quais necessitamos de obter aproximações não de uma mas de uma enorme quantidade de raízes. Basta pensar como seria o caso de no nosso modelo de catenária fazermos variar L, por exemplo, de $L = 90\,m$ a $L = 110\,m$, por acréscimos de $1\,mm$. Para cada valor de L deveríamos determinar a correspondente raiz de $f(a) = 0$ pelo método da bissecção. Se de cada vez realizarmos 25 iterações, no final teríamos efectuado $25 \times 20\,001 = 500\,025$ iterações, o que é manifestamente indesejável.

Tal circunstância sugere a obrigatoriedade de conhecermos algoritmos alternativos que sejam, por um lado de convergência rápida e, por outro, económicos do ponto de vista do número de operações elementares usadas pelo algoritmo, além de numericamente estáveis quando aplicados a um determinado problema. ⧫

2.1.4 Método do ponto fixo

O chamado método do ponto fixo em $\mathbb{R}$, que estudaremos neste parágrafo, é relevante tanto do ponto de vista teórico — trata-se de um método generalizável a espaços mais gerais do que $\mathbb{R}$ — como do ponto de vista computacional pois, frequentemente, este método impõe-se naturalmente a partir de um dado problema concreto. Por exemplo, o método será usado aqui para obtermos aproximações de raízes de uma equação. Mais tarde, no Capítulo 6, veremos que este método pode ser útil nomeadamente no contexto dos chamados métodos *implícitos* para aproximar a solução de uma equação diferencial em que é dado um valor inicial.

Comecemos por definir o conceito de *ponto fixo* e estudar alguns exemplos de motivação.

Definição 2.2. (Ponto fixo)
Seja g uma função real, definida num certo intervalo $[a, b] \subset \mathbb{R}$. O número $z \in [a, b]$ diz-se um ponto fixo de g se $g(z) = z$.

Dada uma função g, determinar os seus pontos fixos equivale a calcular as raízes da equação $g(x) - x = 0$, ou, dito de outra forma, calcular os zeros da função $f(x) = g(x) - x$. Inversamente, se for dada uma equação $f(x) = 0$, calcular as raízes dessa equação equivale a determinar os pontos fixos de uma função g de modo que a equação $g(x) = x$ seja algebricamente equivalente à equação $f(x) = 0$.

Exemplo 2.7. *Pretende-se estudar a existência e localização de pontos fixos reais das seguintes funções iteradoras:*

(a) $g(x) = \alpha\, x + \beta, \quad com \quad \alpha \neq 1, \quad \alpha, \beta \in \mathbb{R}$.
(b) $g(x) = x^2 + 1.$
(c) $g(x) = x^2.$
(d) $g(x) = \cos(x)$.

a) O ponto fixo de g satisfaz a igualdade $\alpha\, z + \beta = z$, ou seja $z = \dfrac{\beta}{1-\alpha}$. Por exemplo, se for $\alpha = 2$ e $\beta = -3$, obtém-se $z = 3$ (ver Fig. 2.4).

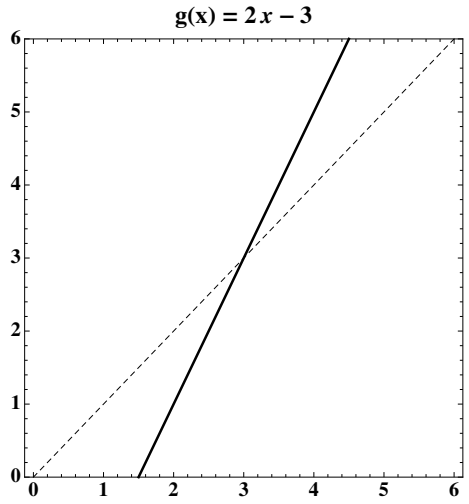


Figura 2.4: Exemplo 2.7 (a).

b) Seja

$$g(x) = x^2 + 1.$$

Neste caso, a equação a ser satisfeita pelos pontos fixos é $z^2 + 1 = z$. Por conseguinte, temos $z = \dfrac{1}{2} \pm \sqrt{\dfrac{1}{2^2} - 1}$, ou seja, não existem pontos fixos reais (ver Fig. 2.1.4).

c)

$$g(x) = x^2.$$

A equação a resolver é $z^2 = z$. Logo, existem dois pontos fixos, $z_1 = 0$ e $z_2 = 1$ (ver Fig. 2.6).

d)

$$g(x) = \cos(x).$$

Embora não seja possível determinar analiticamente o ponto fixo desta função, é fácil verificar que ela tem um ponto fixo (único) no intervalo $[0, 1]$. Com efeito,

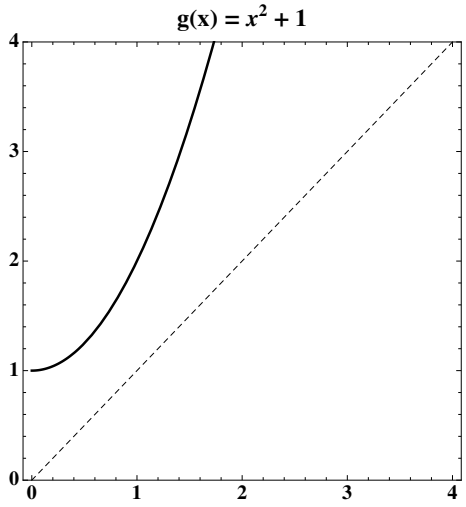


Figura 2.5: Exemplo 2.7 (b).

se definirmos

$$f(x) = \cos(x) - x,$$

verifica-se que $f(0) = 1$ e $f(1) = \cos(1) - 1 < 0$. Logo, sendo a função f contínua, pelo Teorema 2.1 (pág. 31), existe pelo menos um zero z em $]0, 1[$. Nesse ponto verifica-se $\cos(z) = z$, pelo que z é um ponto fixo de g.

Por outro lado, f é uma função continuamente diferenciável e a sua derivada, $f'(x) = -sen(x) - 1$, é negativa em $[0, 1]$. Logo, pelo Teorema 2.2, a função f possui uma única raiz neste intervalo, que é também o único ponto fixo de g (ver Fig. 2.7). ♦

Exemplo 2.8. *Consideremos de novo a equação $e^x - x^2 - 2\,x = 0.5$ (ver Exemplo 2.3, pág. 32).*

A equação pode ser rescrita de várias formas, todas elas equivalentes,

$$\frac{e^x - x^2 - 0.5}{2} = x \tag{2.12}$$

$$\sqrt{e^x - 2\,x - 0.5} = x \tag{2.13}$$

$$\ln(x^2 + 2\,x + 0.5) = x. \tag{2.14}$$

No caso da equação (2.12), as raízes da equação inicial são vistas como os pontos fixos da função $g_1(x) = \dfrac{e^x - x^2 - 0.5}{2}$.

Em relação à equação (2.13), ela remete-nos para os pontos fixos de $g_2(x) = \sqrt{e^x - 2\,x - 0.5}$. Note-se que, neste caso, as equações só são equivalentes para

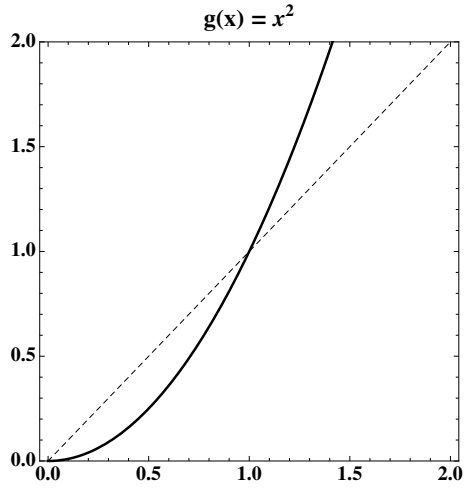


Figura 2.6: Exemplo 2.7 (c).

valores positivos de x (pois a função g_2 toma apenas valores positivos). Em particular, a raiz z_1 sendo negativa não é ponto fixo de g_2 .

Da equação (2.14), concluimos que as raízes da equação inicial são pontos fixos da função $g_3(x) = \ln(x^2 + 2\,x + 0.5)$. Neste caso, a equivalência também não é válida para qualquer valor de x, já que o domínio da função g_3 só inclui os valores de x para os quais $x^2 + 2\,x + 0.5 > 0$. Das raízes da equação inicial apenas z_2 e z_3 satisfazem esta condição. Logo, z_2 e z_3 são também pontos fixos de g_3, enquanto z_1 não o é. ♦

O Exemplo 2.8 mostra-nos que as raízes de uma dada equação $f(x) = 0$ podem ser tratadas como pontos fixos de diferentes funções. Destas funções umas poderão ser úteis para obtermos aproximações numéricas de um determinado ponto fixo, enquanto outras poderão não servir para essa finalidade. Precisamos de saber escolher os métodos numéricos apropriados ao cálculo aproximado desses pontos fixos (ou seja, das raízes de equações equivalentes).

2.1.5 Sucessões numéricas geradas por funções iteradoras

Dada uma função real g, com domínio num certo intervalo $[a, b]$, e um número x_0, tal que $x_0 \in [a, b]$, é possível gerar uma sucessão de números reais $(x_k)_{k\geq 0}$ do seguinte modo:

$$x_{k+1} = g(x_k), \qquad k = 0, 1, ... \tag{2.15}$$

Uma tal sucessão dir-se-á gerada pela função g, ou simplesmente *sucessão gerada por* g.

Se a imagem do intervalo [a,b] estiver contida no próprio intervalo, então a relação (2.15) permite-nos definir uma sucessão infinita de elementos do conjunto considerado. Neste caso, chamaremos a g a função iteradora e aos termos x_k da

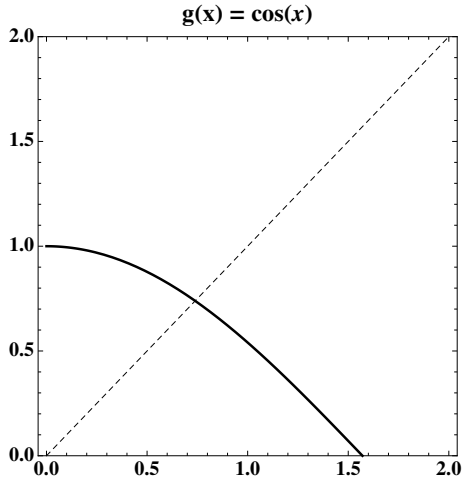


Figura 2.7: Exemplo 2.7 (d).

sucessão as *iteradas.* Veremos como as sucessões geradas desse modo podem ser utilizadas para aproximar as raízes de uma equação dada.

Exemplo 2.9. *Seja*

$$g(x) = x^2.$$

O domínio da função iteradora g *é* $\mathbb{R}$ *(ver Figura. 2.6), e a imagem do intervalo* $[0, 1]$ *por esta função é o próprio intervalo.*

Se tomarmos $x_0 = 0$*, a função* g *gera uma sucessão constante* $\{0, 0, 0, ..\}$.
Se considerarmos $0 < x_0 < 1$*, a sucessão gerada é* $\{x_0, x_0^2, x_0^4, ...\}$ *convergindo para* $x = 0$ *(um dos pontos fixos de* g*).*

Caso se inicie o processo com $x_0 = 1$*, a sucessão das iteradas é de novo constante* $\{1, 1, 1, ..\}$ *(sendo que* $x = 1$ *também é um ponto fixo de* g*).*
Se tomarmos $x_0 > 1$*, a sucessão vai ser divergente (pois tende para infinito).* ♦

O Exemplo 2.9 sugere-nos que quando a sucessão gerada por uma função g converge, o seu limite é um ponto fixo da função g. De facto, assim é:

Teorema 2.4. Seja $(x_n)_{n\geq n_0}$ uma sucessão gerada pela função g, convergindo para um certo limite z. Se g for contínua em z, então z é ponto fixo de g.

Demonstração. Uma vez que $z = \lim_{n\to\infty} x_n$, temos

$$z = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) \ .$$

Da continuidade de g em z resulta que $\lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n) = g(z)$. Obtemos assim que $z = g(z)$, como se pretendia demonstrar. □

Exemplo 2.10. *Considere a sucessão gerada pela função*

$$g(x) = sen(x), \quad com \quad x_0 = 1 \ .$$

Prove que esta sucessão converge. Qual é o seu limite?

Para provar que a sucessão converge basta provar que ela é monótona e limitada.

Note-se que, sendo $0 < x < 1$, temos $0 < sen(x) < x$. Assim,

(i) Todos os termos da sucessão considerada pertencem ao intervalo $[0, 1]$.
(ii) A sucessão é motótona decrescente, visto que $x_{k+1} = sen(x_k) < x_k$. Por conseguinte a sucessão é monótona e limitada, logo é convergente.

De acordo com o Teorema 2.4, a sucessão considerada, sendo convergente, deve convergir para um ponto fixo da função iteradora. O único ponto fixo da função $g(x) = sen(x)$ é $z = 0$, logo é para este ponto que a sucessão de iteradas converge.

♦

2.1.6 Teorema do ponto fixo

O Teorema 2.4 afirma que uma sucessão gerada por uma função iteradora g, a convergir, converge para um ponto fixo daquela função. Fica por responder a questão: sob que condições essa sucessão converge? A resposta a esta questão é dada por um teorema fundamental da Análise, o *teorema do ponto fixo.*

Embora o teorema do ponto fixo possa ser formulado num contexto mais vasto, por agora limitar-nos-emos ao caso em que g é uma função de uma variável real.

Teorema 2.5. (Teorema do ponto fixo)
Seja g uma função real de variável real e $[a, b]$ um intervalo fechado. Se são verificadas as condições:
1)

$$g([a, b]) \subset [a, b] \ .$$

2) A função g é continuamente diferenciável em $[a, b]$.
3)

$$\max_{x \in [a,b]} |g'(x)| = L < 1 \ .$$

Então,

(i) A função g tem um único ponto fixo z em $[a, b]$.
(ii) Se $x_0 \in [a, b]$, a sucessão gerada pela função g converge para o ponto fixo z.

Demonstração. (i) Para demonstrar a existência de pelo menos um ponto fixo, defina-se a função $h(x) = g(x) - x$. Esta função é obviamente contínua em $[a, b]$.

Se $g(a) = a$ (resp. $g(b) = b$), teremos que a (resp. b) é ponto fixo de g. Caso contrário, de acordo com a condição 1), a função h satisfaz

$$h(a) = g(a) - a > 0 \quad \text{e} \quad h(b) = g(b) - b < 0 \ .$$

Assim, pelo Teorema de Bolzano, pág. 31, existe pelo menos um ponto $z \in [a, b]$, tal que $h(z) = 0$, ou seja, $g(z) = z$. Logo, z é ponto fixo de g.

Para demonstrar a unicidade, suponhamos que em $[a, b]$ existem dois pontos fixos distintos $z_1 \neq z_2$. Por definição de ponto fixo temos $g(z_1) = z_1$ e $g(z_2) = z_2$. Logo, $|g(z_1) - g(z_2)| = |z_1 - z_2|$. Por outro lado, usando o Teorema de Lagrange 2.3, pág. 32, e a condição 3), temos

$$|g(z_1) - g(z_2)| \leq \max_{x \in [a,b]} |g'(x)||z_1 - z_2| = L|z_1 - z_2| \ .$$

Donde a desigualdade

$$|z_1 - z_2| \leq L|z_1 - z_2|,$$

ou seja,

$$|z_1 - z_2|(1 - L) \leq 0 \ . \tag{2.16}$$

Mas, de acordo com a condição 3), temos $L < 1$. Logo, da desigualdade (2.16) resulta que $|z_1 - z_2| = 0$, o que contradiz a hipótese de z_1 e z_2 serem distintos. Desta contradição conclui-se a unicidade do ponto fixo.

(ii) Para demonstrar a segunda afirmação, considere-se x_0 um ponto arbitrário de $[a, b]$. Pela condição 1), temos que $x_1 = g(x_0)$ também pertence ao intervalo $[a, b]$. Do mesmo modo se conclui que todos os elementos da sucessão, gerada pela função g, pertencem àquele intervalo.
Vamos agora provar que esta sucessão converge para o ponto fixo z. Pela condição 3), temos

$$|x_n - z| = |g(x_{n-1}) - g(z)| \leq L\,|x_{n-1} - z| \ . \tag{2.17}$$

Aplicando n vezes a desigualdade (2.17), conclui-se que

$$|x_n - z| \leq L^n |x_0 - z| \ . \tag{2.18}$$

Como $L < 1$, da desigualdade (2.18) resulta que $|x_n - z| \to 0$, quando $n \to \infty$ (qualquer que seja $x_0 \in [a, b]$), ou seja, a sucessão $(x_n)_{n \geq 0}$ tende para o ponto fixo z. □

Método do ponto fixo

O teorema do ponto fixo não só garante a existência de um único ponto fixo z da função g num dado intervalo, como indica um método para obter aproximações desse ponto.

Na realidade, se tomarmos qualquer ponto inicial x_0 dentro do intervalo $[a, b]$ e construirmos a sucessão gerada pela função g, de acordo com o teorema do ponto fixo essa sucessão converge para z. O método baseado nesta construção chama-se *método do ponto fixo*.

O método do ponto fixo permite-nos, dada uma função iteradora g e um intervalo $[a, b]$ (satisfazendo as condições (1)-(3) do Teorema 2.5), obter uma aproximação tão precisa quanto quisermos do ponto fixo de g em $[a, b]$.

O algoritmo é extremamente simples:

1. Escolher um ponto $x_0 \in [a, b]$.
2. Calcular cada nova iterada usando a fórmula $x_n = g(x_{n-1})$, $n = 1, 2, \ldots$.
3. Parar quando se obtiver uma aproximação aceitável (critérios de paragem do algoritmo serão abordados adiante).

2.1.7 Estimativas do erro

Para efeitos práticos, interessa-nos não só saber as condições em que um método converge mas também majorar e estimar o erro das aproximações obtidas. No caso do método do ponto fixo, majorações de erro podem obter-se mediante aplicação do seguinte teorema.

Teorema 2.6. Nas condições do Teorema 2.5 são válidas as seguintes estimativas de erro:

$$|x_n - z| \leq L^n \, |x_0 - z| \qquad \text{(majoração } a \textit{ priori}\text{)} \tag{2.19}$$

$$|x_n - z| \leq \frac{L^n}{1 - L} \, |x_1 - x_0| \qquad \text{(majoração } a \textit{ priori}\text{)} \tag{2.20}$$

$$|x_n - z| \leq \frac{L}{1 - L} \, |x_n - x_{n-1}| \qquad n \geq 1, \quad (a \textit{ posteriori}) \tag{2.21}$$

onde x_{n-1} e x_n são duas iteradas consecutivas do método do ponto fixo, e

$$L = \max_{x \in [a,b]} |g'(x)|.$$

Demonstração. A fórmula (2.19) já foi obtida na demonstração do teorema do ponto fixo (ver (2.18), pág. 48).

Quanto à desigualdade (2.21), comecemos por observar que

$$|x_{n-1} - z| = |z - x_{n-1}| \leq |z - x_n| + |x_n - x_{n-1}|. \tag{2.22}$$

Por outro lado, de acordo com (2.17), temos

$$|x_n - z| \leq L\,|x_{n-1} - z|,$$

e portanto

$$|x_{n-1} - z|\,(1 - L) \leq |x_n - x_{n-1}|\ . \tag{2.23}$$

Observando que $1 - L > 0$ (atendendo à condição 3) do Teorema 2.5) podem dividir-se por este valor ambos os membros da desigualdade (2.23), obtendo-se

$$|x_{n-1} - z| \leq \frac{1}{1-L}\,|x_n - x_{n-1}|\ . \tag{2.24}$$

Finalmente, das desigualdades (2.24) e (2.17) resulta a estimativa (2.21).

A expressão (2.20) resulta de (2.21). Com efeito, para $n = 1$, tem-se

$$|z - x_1| \leq \frac{L}{1-L}\,|x_1 - x_0|\ . \tag{2.25}$$

Para $n = 2$, atendendo a (2.19), é válida a desigualdade

$$|z - x_2| \leq L\,|z - x_1|\ .$$

Levando em consideração (2.25), resulta

$$|z - x_2| \leq \frac{L^2}{1-L}\,|x_1 - x_0|\ .$$

De modo análogo, conclui-se por indução (2.20). □

As estimativas de erro discutidas no parágrafo 2.1.2, pág. 33, são aplicáveis para métodos de ponto fixo satisfazendo as condições do Teorema 2.1.6. Com efeito, fazendo

$$f(x) = g(x) - x,$$

tem-se que $g(z) = z$ se e só se $f(z) = 0$. Além disso,

$$f'(x) = g'(x) - 1\ .$$

Atendendo a que no intervalo $I = [a, b]$, por hipótese $0 \leq |g'(x)| < 1$, resulta que $f'(x) \neq 0,\ \forall x \in I$. Logo,

$$m = \min_{a \leq x \leq b} |f'(x)| = \min_{a \leq x \leq b} |g'(x) - 1| > 0, \tag{2.26}$$

pelo que a majoração de erro (2.6) de cada aproximação x_k de z obtida por aplicação da função iteradora g, é da forma

$$|e_k| = |z - x_k| \leq \frac{|g(x_k) - x_k|}{m} = \frac{|x_{k+1} - x_k|}{m}, \tag{2.27}$$

e a respectiva estimativa de erro (2.5), escreve-se

$$\bar{e}_k = -\frac{g(x_k) - x_k}{g'(x_k) - 1} = -\frac{x_{k+1} - x_k}{g'(x_k) - 1} = \frac{x_{k+1} - x_k}{1 - g'(x_k)}\ . \tag{2.28}$$

Exemplo 2.11. *Considere a equação* $cos(x) - 2\,x = 0$.

(a) Com base no teorema do ponto fixo mostre que esta equação tem uma única raiz no intervalo $[0.4, 0.5]$*, e que o método do ponto fixo converge para essa raiz.*

(b) Tomando como aproximação inicial $x_0 = 0.4$*, calcule as duas primeiras iteradas do método.*

(c) Obtenha uma estimativa do erro da aproximação x_2 *calculada na alínea anterior.*

(d) Nas condições da alínea (c), quantas iterações é necessário efectuar para garantir que o erro absoluto da aproximação obtida seja inferior a 0.001*?*

(a) Comecemos por observar que qualquer raiz da equação dada é um ponto fixo de $g(x) = \dfrac{\cos(x)}{2}$.

Mostremos agora que a função g satisfaz as condições do teorema do ponto fixo no intervalo referido. Para o efeito, comecemos por calcular as imagens dos extremos do intervalo,

$$\begin{aligned} g(0.4) = \cos(0.4)/2 = 0.46053 \quad &\in [0.4, 0.5] \\ g(0.5) = \cos(0.5)/2 = 0.43879 \quad &\in [0.4, 0.5]. \end{aligned}$$

Por outro lado, a função g é decrescente em $[0.4, 0.5]$ (pois $g'(x) = -\sin(x)/2$ é negativa naquele intervalo), donde se conclui que $g([0.4, 0.5]) \subset [0.4, 0.5]$.

A função g é continuamente diferenciável em $\mathbb{R}$ e, em particular, no intervalo considerado. Tem-se,

$$L = \max_{x\in[0.4,0.5]} |g'(x)| = \max_{x\in[0.4,0.5]} \frac{|\sin x|}{2} = \frac{\sin(0.5)}{2} = 0.2397 < 1 \ .$$

Todas as condições do teorema do ponto fixo estão satisfeitas, pelo que o método do ponto fixo com a função iteradora $g(x) = \cos(x)/2$ converge para o ponto fixo.

(b) Tomando como aproximação inicial $x_0 = 0.4$, as duas primeiras aproximações iniciais são

$$\begin{aligned} x_1 &= g(x_0) = 0.46053 \\ x_2 &= g(x_1) = 0.44791 \ . \end{aligned}$$

(c) Usando a fórmula (2.21), obtém-se

$$|z - x_2| \leq \frac{L}{1-L}|x_2 - x_1| = \frac{0.2397}{1 - 0.2397}|0.44791 - 0.46053| = 0.00397.$$

(d) Para responder a esta questão podemos aplicar a estimativa *a priori* (2.19). De acordo com esta estimativa, temos

$$|x_n - z| \leq L^n\,|x_0 - z| \leq 0.2397^n\,|0.5 - 0.4| = 0.1 \times 0.2397^n, \qquad n \geq 1 \ .$$

Logo, para garantir que o erro absoluto da n-ésima iterada é inferior a uma certa tolerância ϵ, basta escolher n de tal modo que $0.2397^n < 10\,\epsilon$. Desta inequação, resulta

$$n > \frac{\ln(10\,\epsilon)}{\ln 0.2397} \simeq 3.22, \quad \text{para} \quad \epsilon = 10^{-3}.$$

Donde se conclui que bastam 4 iterações para satisfazer a tolerância de erro exigida.
Na realidade apenas são necessárias três iterações para satisfazer a referida tolerância de erro. Para $z = 0.450\,183\,611\,295$, efectuando cálculos com arredondameno simétrico para 12 casas decimais, pode verificar-se que $g(z) = z$, pelo que tomaremos o valor indicado como ponto fixo da função iteradora g considerada. Na tabela a seguir mostram-se os resultados da aplicação das estimativas de erro (2.28) e majorações de erro (2.27), com m dado por (2.26), iniciando o processo com $x_0 = 0.4$, e $m = \min_{0.4\le x\le 0.5} |g'(x) - 1| = (2 + \sin(0.4))/2 \simeq 1.19470917$.

k	x_k	$e_k = z - x_k$	$\bar{e}_k = \dfrac{g(x_k) - x_k}{g'(x_k) - 1}$	$\lvert e_k\rvert \le \dfrac{\lvert g(x_k) - x_k\rvert}{m}$
0	0.4	0.0501836113	0.0506654661	0.0506654661
1	0.460530497001	−0.0103468857	−0.0103272353	0.0105649710
2	0.447908429155	0.00227518214	0.00227614062	0.00231773354
3	0.450677446670	−0.000493835	−0.000493790	0.000503328999

Note-se a mudança de sinal das (boas) estimativas de erro $\bar{e}_k$, indicando que as sucessivas iteradas se localizam ora à esquerda ora à direira do ponto fixo z. A iterada x_3 satisfaz a desigualdade $|z - x_3| < 10^{-3}$. ♦

2.1.8 Classificação de pontos fixos

De acordo com o teorema do ponto fixo, a convergência das sucessões geradas por uma certa função g num intervalo $[a, b]$ depende do comportamento da sua derivada g' nesse intervalo. Isto leva-nos a classificar os pontos fixos z de uma função g de acordo com o valor de $g'(z)$.

Neste parágrafo iremos assumir que a função $g \in C^1$ (ou seja, g e g' são funções contínuas), pelo menos numa vizinhança de cada ponto fixo de g, caso em que diremos ser g uma função iteradora *regular*.

Definição 2.3. Um ponto fixo z, de uma função iteradora regular g, diz-se:
Atractor, se $0 < |g'(z)| < 1$;
Superatractor, se $g'(z) = 0$;
Repulsor, se $|g'(z)| > 1$;
Neutro, se $|g'(z)| = 1$.

De facto, se $|g'(z)| < 1$ e g' é contínua em z, então existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$ tal que $\max_{x\in V_\epsilon(z)} |g'(z)| = L < 1$. Por outro lado, se $x \in V_\epsilon(z)$,

temos

$$|g(x) - g(z)| \leq L\,|x - z| < |x - z| < \epsilon,$$

ou seja, $g(x)$ também pertence a $V_\epsilon(z)$.

Logo, se o intervalo $[a, b]$ estiver contido em $V_\epsilon(z)$, nesse intervalo a função g satisfaz as condições do teorema do ponto fixo.

Concluímos portanto que, *se z for um ponto fixo atractor, então existe uma vizinhança $V_\epsilon(z)$ tal que, se $x_0 \in V_\epsilon(z)$, então a sucessão gerada por g converge para z.*

No caso $g'(z) > 1$, é fácil verificar que nenhuma sucessão gerada pela função g converge para z (excepto a sucessão constante $z, z, \ldots$, ou qualquer sucessão da forma $\ldots, x, z, z, \ldots$, onde x é tal que $g(x) = z$).

Com efeito, se z é um ponto fixo repulsor, existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$ tal que $|g'(z)| > 1, \forall x \in V_\epsilon(z)$. Assim, seja x_k um termo de uma sucessão gerada pela função g e suponhamos que $x_k \in V_\epsilon(z)$, com $x_k \neq z$. Tem-se,

$$|x_{k+1} - z| = |g(x_k) - g(z)| \geq \min_{x \in V_\epsilon(z)} |g'(x)||x_k - z| > |x_k - z|.$$

Logo, x_{k+1} está mais distante de z do que x_k. Se o ponto x_{k+1} também pertencer a $V_\epsilon(z)$, o mesmo raciocínio aplica-se a esse ponto, e vemos que a sucessão se afasta de z.
A única possibilidade de uma sucessão não constante convergir para z, sendo z repulsor, é o caso dessa sucessão conter um ponto x (não pertencente à vizinhança referida), tal que $g(x) = z$.

Quando o ponto fixo é neutro, isto é, $|g'(z)| = 1$, existem sucessões geradas pela função g que convergem para z e outras que não convergem (mesmo que x_0 esteja próximo do ponto fixo z), justificando-se assim a designação dada a um ponto fixo desta natureza.

O caso do ponto fixo superatractor merece atenção particular, pois o facto de se ter $g'(z) = 0$, indica que o método iterativo correspondente convergirá muito rapidamente para o ponto fixo, como teremos oportunidade de discutir mais adiante.

Exemplo 2.12. *Consideremos a função*

$$g(x) = k\,x(1 - x), \quad \textit{onde} \quad k > 0\ .$$

Esta função é conhecida como "função logística". Tal função iteradora aparece no contexto de modelos matemáticos da Ecologia.

Vamos determinar os pontos fixos da equação $x = g(x)$ e classificá-los segundo a Definição 2.3.

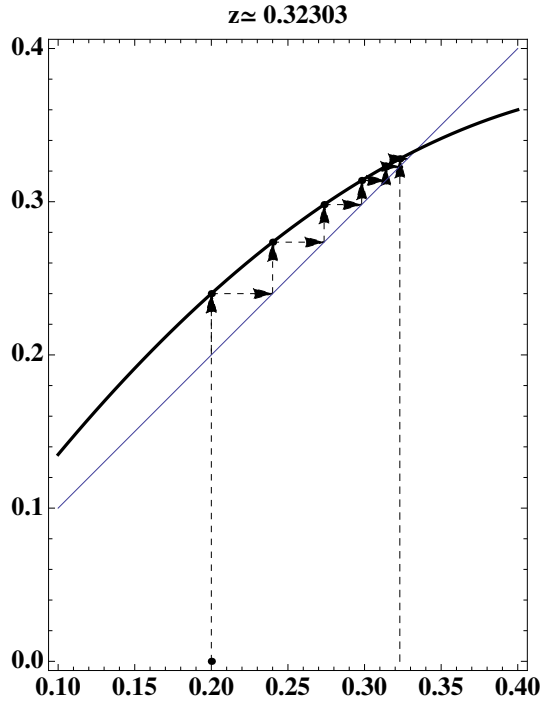


Figura 2.8: Iterações da função $g(x) = 1.5\,x\,(1-x)$, com $x_0 = 0.2$.

Para determinarmos os pontos fixos da função g, para um certo valor de k dado, resolva-se a equação

$$k\,z(1-z) = z\ . \tag{2.29}$$

É fácil verificar que esta equação possui duas raízes, $z_1 = 0$ e $z_2 = 1 - 1/k$.

Vejamos como classificar os pontos fixos em causa.
Consideremos, por exemplo, o caso $k = 1.5$. Os dois pontos fixos de g são $z_1 = 0$ e $z_2 = 1/3$. Para os classificarmos, observemos que $g'(x) = 1.5 - 3\,x$. Logo $g'(0) = 1.5$ e $g'(1/3) = 1.5 - 1 = 0.5$, ou seja, z_1 é ponto fixo *repulsor*, e z_2 é *atractor*.

Por conseguinte:

a) Nenhuma sucessão gerada pela função g poderá convergir para 0 (excepto a sucessão constante, igual a 0, ou a sucessão $1, 0, 0, \ldots$).

b) Se x_0 for suficientemente próximo de $1/3$, a sucessão gerada por g converge para $z_2 = 1/3$. Mais precisamente, pode provar-se que, se $0 < x_0 < 1$, a sucessão $(x_k)_{k\geq 0}$ converge para z_2. As Figuras 2.8 e 2.9 ilustram esta afirmação.

♦

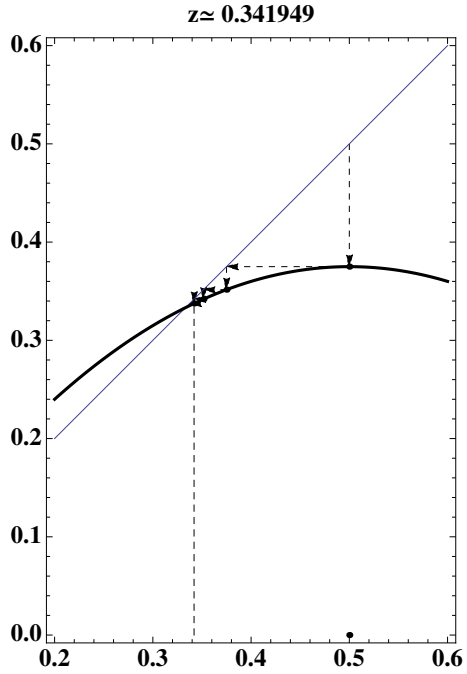


Figura 2.9: Iterações da função $g(x) = 1.5\,x\,(1-x)$, com $x_0 = 0.5$.

Exemplo 2.13. *Vejamos que a função iteradora*

$$g(x) = x^2 + x$$

possui um ponto fixo neutro.

A função iteradora g tem um ponto fixo (único) $z = 0$. Visto que

$$g'(z) = 2\,z + 1 = 1,$$

este ponto fixo é *neutro*.

Vejamos agora qual é o comportamento das sucessões geradas por esta função. Considerando $x_0 = 0.12$, as duas primeiras iteradas são

$$\begin{aligned} x_1 &= x_0^2 + x_0 = 0.1344 \\ x_2 &= x_1^2 + x_1 = 0.152463\ . \end{aligned}$$

É fácil verificar que, neste caso, a sucessão é crescente e tende para $+\infty$. Se escolhermos como ponto inicial $x_0 = -0.12$, obtém-se

$$\begin{aligned} x_1 &= x_0^2 + x_0 = -0.1056 \\ x_2 &= x_1^2 + x_1 = -0.0945\ . \end{aligned}$$

A sucessão é crescente e converge para o ponto fixo $z = 0$. As figuras 2.10 e 2.11 ilustram este exemplo. ♦

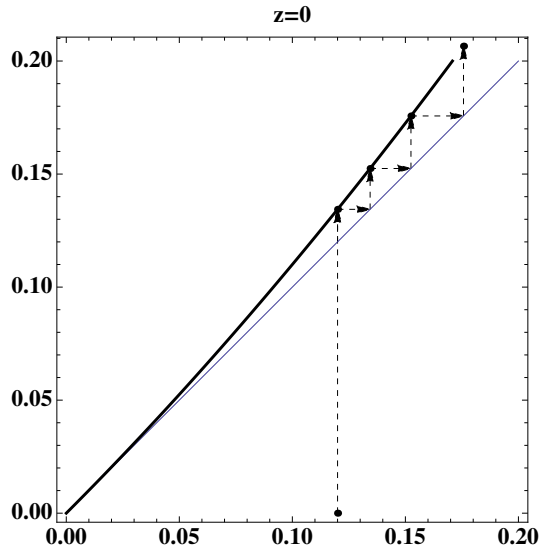


Figura 2.10: Iterações da função $g(x) = x^2 + x$, com $x_0 = 0.12$.

Exemplo 2.14. *Na pág. 30 foi definida uma função $\phi(h)$, a partir da qual se resolve a equação $\phi(h) = 0$. A partir dessa equação obtém-se a função*

$$g(h) = \phi(h) + h,$$

definida no intervalo $[0, 50]$, a qual poderá servir para determinar a altura h no problema da catenária tratado no Exemplo 2.2, pág. 29, onde se discutiu o problema da catenária. Na Figura 2.12 encontra o gráfico de g no intervalo considerado.

A função g possui um único ponto fixo em $[0, 50]$. Se escolhermos uma estimativa inicial $h_0 \in [0, 50]$, poderemos usar o método de ponto fixo, com a função iteradora g, para determinar esse ponto fixo?

A observação do gráfico é suficiente para concluirmos que existe um único ponto fixo da função g (próximo de $h = 30$), mas deveremos usar com reservas o método de ponto fixo com tal função iteradora. De facto, $g'(z) \simeq 1$, ou seja, o ponto fixo (embora atractor) conduzirá necessariamente a um processo de convergência lenta. Veremos adiante, no parágrafo 2.4, como contornar esse problema. ♦

2.1.9 Observações sobre monotonia das iteradas

Suponhamos que z é um ponto fixo atractor ou superatractor da função g. Como se referiu no parágrafo anterior é satisfeita a condição $|g'(z)| < 1$, isto é, $-1 < g'(z) < 1$. Neste caso, qualquer sucessão gerada pela função g, com x_0 suficientemente próximo de z, converge para z.

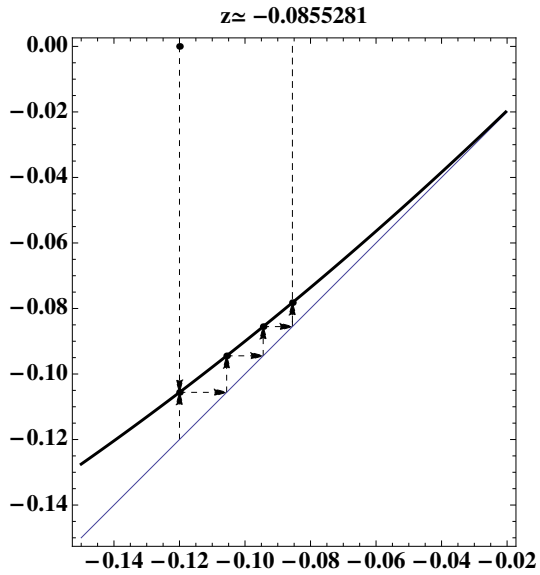


Figura 2.11: Iterações da função $g(x) = x^2 + x$, com $x_0 = -0.12$.

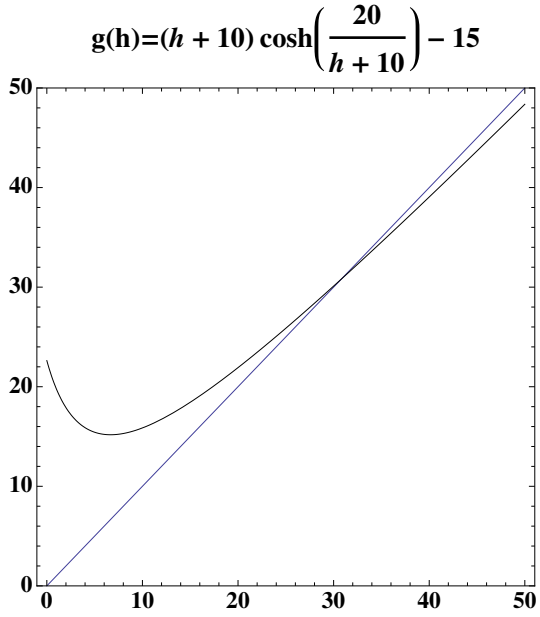


Figura 2.12: Função $g(h)$ do Exemplo 2.14.

Neste parágrafo, vamos investigar em que condições essa sucessão é monótona (crescente ou decrescente). Tal como antes, admitimos que g é continuamente diferenciável numa vizinhança de z.

Caso 1. Suponhamos que

$$0 \leq g'(z) < 1 \ .$$

Da continuidade da derivada de g, resulta que existe uma vizinhança $V_\epsilon(z) = (z-\epsilon, z+\epsilon)$, tal que, se $x \in V_\epsilon(z)$ então $0 < g'(x) < 1$.

Suponhamos que x_k é um termo de uma sucessão gerada pela função g, tal que $x_k \in V_\epsilon(z)$. Para sermos mais específicos, admitamos que $z < x_k < z + \epsilon$. Nesse caso, uma vez que $x_{k+1} = g(x_k)$, aplicando o Teorema de Lagrange, pág. 32, existe um ponto ξ_k, com $z \leq \xi_k \leq x_k$, tal que

$$x_{k+1} - z = g(x_k) - g(z) = g'(\xi_k)(x_k - z) \ . \tag{2.30}$$

Por construção, temos $x_k - z > 0$ e $g'(\xi_k) > 0$. Logo, $x_{k+1} > z$. Concluímos portanto que se $x_k > z$ então também $x_{k+1} > z$.

Por outro lado, uma vez que z é um ponto atractor (é verdade que $0 < g'(\xi_k) < 1$), pelo que o ponto x_{k+1} deve estar mais próximo de z do que x_k, donde se conclui que $x_{k+1} < x_k$. Como o mesmo raciocínio se aplica a todas as iteradas subsequentes, podemos dizer que, neste caso, a sucessão $(x_n)_{n\geq k}$ é *decrescente* (pelo menos, a partir da ordem k). Esta situação é ilustrada, por exemplo, no gráfico da Figura 2.9.

Analogamente, se tivermos $x_k < z$, podemos concluir que $x_{k+1} > x_k$, pelo que a sucessão das iteradas será *crescente* (ver Figuras 2.8 e 2.11). Em qualquer dos casos, as respectivas sucessões das iteradas são *monótonas*.

Caso 2. Suponhamos agora que

$$-1 < g'(z) < 0 \ .$$

Da continuidade da derivada de g, resulta que existe uma vizinhança $V_\epsilon(z) = (z-\epsilon, z+\epsilon)$, tal que : se $x \in V_\epsilon(z)$ então $-1 < g'(x) < 0$.

Admitindo que x_k pertence a essa vizinhança, a igualdade (2.30) é aplicável. Neste caso, supondo que $x_k > z$, dessa igualdade resulta que $x_{k+1} < z$ (uma vez que $g'(\xi_k) < 0$). Se aplicarmos o mesmo raciocínio às iteradas seguintes, concluímos que $x_{k+2} > z$, $x_{k+3} < z$, etc.

Se, pelo contrário, tivermos $x_k < z$, então $x_{k+1} > z$, $x_{k+2} < z$, etc. Ou seja, neste caso, as iteradas vão ser alternadamente maiores ou menores que z (uma sucessão deste tipo diz-se *alternada*).

Caso 3. Se $g'(z) = 0$ (ponto fixo superatractor) é necessária informação suplementar sobre as derivadas de g, de ordem superior, para que se possa decidir algo sobre a monotonia da sucessão das respectivas iteradas.

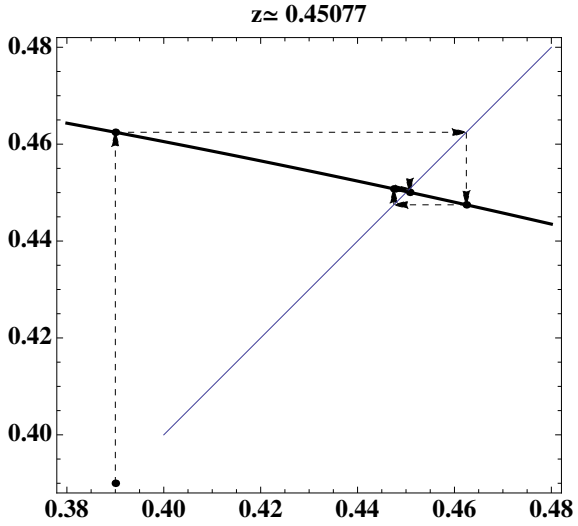


Figura 2.13: Iterações da função $g(x) = \dfrac{\cos(x)}{2}$, com $x_0 = 0.39$.

2.1.10 Sucessões alternadas

Uma propriedade interessante das sucessões alternadas convergentes é que o limite da sucessão está sempre localizado entre dois termos consecutivos, isto é, $x_k < z < x_{k+1}$, ou $x_{k+1} < z < x_k$. Tal facto permite-nos obter um majorante do erro absoluto de x_{k+1}, além daqueles que já obtivemos. Tem-se

$$|x_{k+1} - z| < |x_{k+1} - x_k| \ . \tag{2.31}$$

A sucessão das iteradas do Exemplo 2.11, pág. 51, em que $g'(z) < 0$, é um exemplo de uma sucessão alternada. Na Figura 2.13 estão representados graficamente alguns termos desta sucessão.

2.1.11 Divergência do método do ponto fixo

O estudo de pontos fixos repulsores iniciado no parágrafo 2.1.8, pág. 52, permite-nos formular o seguinte critério de divergência do método do ponto fixo.

Teorema 2.7. Seja g uma função iteradora continuamente diferenciável em $[a, b]$, tal que

$$|g'(x)| > 1, \quad \forall x \in [a, b]$$

e z ponto fixo de g.
Exceptuando a sucessão constante $z, z, \ldots$, ou qualquer sucessão da forma $\ldots, x, z, z, \ldots$, nenhuma sucessão gerada pela funçao g pode convergir no intervalo $[a, b]$.

Demonstração. De acordo com as hipóteses formuladas e com a classificação dos pontos fixos na página 52, se a função g tiver algum ponto fixo em $[a, b]$, esse ponto fixo é repulsor. Por outro lado, se uma sucessão gerada pela função g convergir, ela converge para um ponto fixo de g (Teorema 2.4, pág. 46). Da conjugação destes dois factos resulta a afirmação no enunciado. □

2.2 Ordem de convergência

Um dos conceitos fundamentais da teoria dos métodos iterativos refere-se à sua *ordem de convergência*. Este conceito permite-nos comparar a rapidez com que diferentes métodos convergem e escolher, em cada caso, o método mais rápido.

Representaremos por $(x_n)_{n\geq n_0}$ (n_0 é o índice do primeiro termo da sucessão, geralmente $n_0 = 0$ ou $n_0 = 1$), uma sucessão convergente para z.

Definição 2.4. Diz-se que uma sucessão $(x_n)_{n\geq n_0}$ convergente para z, possui convergência *de ordem* $p > 1$, com $p \in \mathbb{R}$, se existir uma constante $k_\infty > 0$ tal que

$$k_\infty = \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|^p} \ .$$

A constante k_∞ designa-se por *coeficiente assimptótico de convergência*.

No caso particular de $p = 1$, diz-se que a convergência é *linear*. Quando $p > 1$ a convergência diz-se *supralinear*.

Note-se que no caso $p = 1$, o coeficiente $0 < k_\infty < 1$ permite-nos comparar quanto à rapidez de convergência métodos distintos que possuam convergência linear. Com efeito, quanto mais pequeno (mais próximo de 0) for o valor de k_∞, mais rápida será a convergência.

Exemplo 2.15. *Consideremos a sucessão* $(x_n)_{n\geq 0}$*, tal que*

$$x_{n+1} = \frac{x_n}{a}, \quad \textit{para} \quad a > 1, \quad \textit{com} \quad x_0 \in \mathbb{R}.$$

A sucessão converge? E sendo convergente, é de convergência linear ou supralinear?

É fácil verificar que esta sucessão converge para $z = 0$, qualquer que seja $x_0 \in \mathbb{R}$, já que este é o único ponto fixo da função iteradora $g(x) = x/a$. Além disso, este ponto fixo é atractor, visto que $g'(x) = 1/a < 1$, para todo o $x \in \mathbb{R}$.

Verifiquemos que a sucessão possui convergência linear. Para isso, calculemos

$$k_\infty = \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|} = \lim_{n\to\infty} \frac{|x_{n+1}|}{|x_n|} = \frac{1}{a} < 1 \ . \tag{2.32}$$

Concluímos assim que a convergência é linear e o coeficiente assimptótico de convergência é $k_\infty = \dfrac{1}{a}$. A convergência será tanto mais rápida quanto maior for a.

Que conclusões pode tirar deste processo iterativo quando $a = 1$? ⧫

Analisemos agora um exemplo em que a ordem de convergência é superior a um.

Exemplo 2.16. *Considere a sucessão $(x_n)_{n\geq 0}$, tal que*

$$x_{n+1} = b\, x_n^\alpha, \quad \textit{onde} \quad b \neq 0 \quad \textit{e} \quad \alpha > 1, \ \textit{com} \quad |x_0| < |b|^{\frac{-1}{\alpha-1}} \ .$$

Mostre que a sucessão converge para $z = 0$, e estude a sua ordem de convergência.

É fácil verificar que esta sucessão converge para $z = 0$, se x_0 satisfizer a condição indicada. De facto, o ponto $z = 0$ é um ponto fixo *superatractor* para a função iteradora $g(x) = b\, x^\alpha$, visto que $g'(0) = 0$.
Por outro lado, sendo $|x_0| < |b|^{\frac{-1}{\alpha-1}}$, resulta $|x_1| < |x_0|$ e, de um modo geral, teremos que $|x_{n+1}| < |x_n|$, $\forall n \geq 0$. Isto é, a sucessão é decrescente em módulo, pelo que converge para $x = 0$.

Verifiquemos qual a respectiva ordem de convergência. Para o efeito calculemos o limite,

$$\lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|^p} = \lim_{n\to\infty} \frac{|x_{n+1}|}{|x_n|^p} = \lim_{n\to\infty} \frac{|b\, x_n^\alpha|}{|x_n|^p} \ . \tag{2.33}$$

Para que este limite seja finito, deveremos ter $p = \alpha$. Neste caso, $k_\infty = |b|$ e portanto a ordem de convergência é α (convergência supralinear), e o coeficiente assimptótico de convergência vale $|b|$. ⧫

2.2.1 Convergência supralinear

Nos métodos de convergência supralinear (em particular os métodos de Newton e da secante a discutir adiante (pág. 65 e pág. 80), sendo x_k uma aproximação de um zero simples z de uma função real f, o erro $e_k = z - x_k$ pode ser estimado pela diferença $z - x_k \simeq x_{k+1} - x_k$, isto é,

$$\bar{e}_k = x_{k+1} - x_k \ . \tag{2.34}$$

Com efeito, para $z \neq x_k$, de

$$z - x_{k+1} = z - x_k - (x_{k+1} - x_k),$$

obtém-se

$$\frac{z - x_{k+1}}{z - x_k} = 1 - \frac{x_{k+1} - x_k}{z - x_k} \ .$$

Uma vez que por hipótese a sucessão $(x_k)_{k\geq 0}$ converge supralinearmente para z, resulta que

$$0 = \lim_{k\to\infty} \left| \frac{z - x_{k+1}}{z - x_k} \right| = \lim_{k\to\infty} \left| 1 - \frac{x_{k+1} - x_k}{z - x_k} \right| .$$

Consequentemente, por definição de limite de uma sucessão, tem-se

$$\lim_{k\to\infty} \frac{x_{k+1} - x_k}{z - x_k} = 1 . \tag{2.35}$$

A igualdade em (2.35) diz-nos que as sucessões $(x_{k+1} - x_k)_{k\geq 0}$ e $(e_k)_{k\geq 0}$ são *assimptoticamente iguais*, isto é, para k suficientemente grande (2.34) dá uma boa estimativa do erro do termo x_k da sucessão.

Exemplo 2.17.

No Exemplo 2.18, pág. 64, mostramos que a sucessão

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{1}{x_k} \right), \quad k \geq 0$$

converge quadraticamente (ordem $p = 2$) para o número $z = 1$, independentemente da escolha que se fizer do termo inicial $x_0 \in [1, 2]$.

Para $x_0 = 2$, e calculando alguns dos termos seguintes da sucessão, compara-se na tabela abaixo o erro exacto $z - k$, com a estimativa $\bar{e}_k$. Na coluna mais à direita da tabela inscreveram-se os quocientes $(x_{k+2} - x_{k+1})/(x_{k+1} - x_k)^2$, os quais aproximam a constante assimptótica de convergência da sucessão, $k_\infty = 1/2$. Os cálculos foram efectuados usando precisão estendida do sistema *Mathematica*.

A fim de manter a tabela dentro dos limites da página os valores inscritos nas terceira e quarta colunas foram convenientemete truncados. A qualidade de cada uma das aproximações de erro e_k é evidenciada através dos quocientes referidos os quais dão boas estimativas de k_∞ logo a partir de $x_1 = 1.25$.

x_k	$z - x_k$	$\bar{e}_k = x_{k+1} - x_k$	$\dfrac{x_{k+2} - x_{k+1}}{(x_{k+1} - x_k)^2}$
2	-1	-0.75	0.4
1.2500000000000000000	-0.25	-0.225	0.487804878
1.0250000000000000000	-0.025	-0.024695121	0.499847607
1.0003048780487804879	-0.0003048780487804878	-0.000304831	0.499999976

♦

2.2.2 Ordem de convergência de métodos do ponto fixo

A ordem de convergência de um determinado método do ponto fixo depende das propriedades de regularidade da respectiva função iteradora g.

O teorema que se segue diz-nos quais as condições que a função g deve satisfazer para garantir que o método do ponto fixo possua convergência pelo menos linear. Uma vez satisfeitas tais condições, poderemos assegurar que o método possui uma certa ordem de convergência $p \geq 1$, sendo p um certo inteiro positivo.

Teorema 2.8. (Ordem de convergência do método do ponto fixo)
Seja $p \geq 1$, e g uma função de classe C^p em $[a, b]$, satisfazendo as condições do teorema do ponto fixo nesse intervalo, e $z \in [a, b]$ ponto fixo da função iteradora g. Se

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

então:

(1) A função g possui um único ponto fixo z em $[a, b]$.

(2) Se $x_0 \in [a, b]$, a sucessão gerada por g converge para z, com ordem de convergência p.

(3) O coeficente assimptótico de convergência é $k_\infty = \dfrac{|g^{(p)}(z)|}{p!}$.

Demonstração. A primeira afirmação resulta do teorema do ponto fixo, pág. 47.

Resta-nos provar os itens (2) e (3). Para o efeito, considere-se o desenvolvimento de Taylor g, em torno de z,

$$g(x) = g(z) + g'(z)(x - z) + \frac{g''(z)}{2}(x - z)^2 + \cdots + \frac{g^{(p)}(\xi)}{p!}(x - z)^p, \qquad (2.36)$$

onde $\xi \in int(z, x)$[7]. Em particular, se escrevermos a fórmula (2.36) com $x = x_m$, atendendo às hipóteses formuladas, obtém-se

$$g(x_m) = g(z) + \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p, \qquad (2.37)$$

onde $\xi_m \in int(z, x_m)$. Uma vez que $g(z) = z$ e $x_{m+1} = g(x_m)$, da fórmula (2.37) resulta imediatamente

$$x_{m+1} - z = \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p. \qquad (2.38)$$

Dividindo ambos os membros de (2.38) por $(x_m - z)^p$ e tomando o módulo, obtém-se

$$\frac{|x_{m+1} - z|}{|x_m - z|^p} = \frac{|g^{(p)}(\xi_m)|}{p!} \ . \qquad (2.39)$$

[7] A notação $int(z, x)$ significa tratar-se de um intervalo aberto, onde o extremo inferior é o mínimo dos valores z e x, e o extremo superior o máximo desses dois valores.

Calculando o limite quando $m \to \infty$, de (2.39), obtém-se

$$\lim_{m\to\infty} \frac{|x_{m+1} - z|}{|x_m - z|^p} = \frac{|g^{(p)}(z)|}{p!} \ . \tag{2.40}$$

Da igualdade (2.40) resulta imediatamente que a sucessão (x_m) possui ordem de convergência p, e que $k_\infty = \dfrac{|g^{(p)}(z)|}{p!}$. □

Observação. Como caso particular do Teorema 2.7, quando $p = 1$, conclui-se que se g satisfizer as condições do teorema do ponto fixo em $[a, b]$, e se $g'(z) \neq 0$, então qualquer que seja $x_0 \in [a, b]$, a sucessão gerada pela função g converge linearmente para z, e o coeficiente assimptótico de convergência é $k_\infty = |g'(z)|$. Por conseguinte, a convergência será tanto mais rápida quanto mais próximo de 0 for o valor de k_∞.

Exemplo 2.18. *Considere a função iteradora*

$$g(x) = \frac{1}{2}\left(x + \frac{1}{x}\right).$$

(a) Mostre que os pontos fixos de g são $z_1 = 1$ e $z_2 = -1$.
(b) Classifique esses pontos fixos.
(c) Para $x_0 \in [1, 2]$, mostre que a sucessão gerada pela função g converge para $z_1 = 1$, e determine a ordem bem como o coeficiente assimptótico de convergência.

(a) A igualdade $g(z) = z$ é equivalente a

$$g(z) = \frac{1}{2}\left(z + \frac{1}{z}\right) = z \quad \Longrightarrow \ z^2 + 1 = 2\,z^2 \iff z^2 = 1 \ .$$

Assim, os pontos fixos de g são $z_1 = 1$ e $z_2 = -1$.

(b) Visto que $g'(x) = \dfrac{1}{2} - \dfrac{1}{2\,x^2}$, obtém-se $g'(1) = g'(-1) = 0$, ou seja, estes pontos fixos são superatractores.

(c) Mostremos que a função g satisfaz as condições do teorema do ponto fixo em $[1, 2]$.

Já sabemos que $g'(x) = \dfrac{1}{2} - \dfrac{1}{2\,x^2}$, logo a função g é continuamente diferenciável em $[1, 2]$. Além disso, verifica-se facilmente que $g'(x) \geq 0$, para todo $x \in [1, 2]$, pelo que g é crescente em $[1, 2]$.

Para se mostrar que $g([1, 2]) \subset [1, 2]$, basta verificar que $g(1) = 1 \in [1, 2]$ e $g(2) = 5/4 \in [1, 2]$. Por outro lado, temos $\max_{x\in[1,2]} |g'(x)| = |g(2)| = \dfrac{3}{8} < 1$.

Tendo em vista determinarmos a ordem de convergência e o coeficiente assimptótico de convergência da sucessão considerada, vamos aplicar o Teorema 2.7, pág. 59.

Para o efeito, analisemos as derivadas de g. Já sabemos que $g'(1) = 0$. Quanto à segunda derivada, temos $g''(x) = \dfrac{1}{x^3}$. Logo, g'' é contínua em $[1, 2]$, e $g''(1) = 1 \neq 0$. Daqui resulta que o Teorema 2.7 é aplicável, sendo a ordem de convergência $p = 2$.

Quanto ao coeficiente assimptótico de convergência, temos

$$k_\infty = \frac{|g''(1)|}{2} = \frac{1}{2} \, .$$

O valor calculado para o coeficiente assimptótico de convergência, $k_\infty = 0.5$, indica que para n suficientemente grande se tem

$$|z - x_{n+1}| \simeq 0.5 \, |z - x_n|^2.$$

Ou seja, a partir de certa ordem, o erro de cada iterada é aproximadamente igual a 50 % do quadrado do erro da iterada anterior. ⧫

2.3 Método de Newton

Na secção anterior vimos que o método do ponto fixo tem um vasto domínio de aplicação e permite, com frequência, obter boas aproximações de raízes de equações. No entanto, em geral aquele método garante apenas primeira ordem de convergência – ordens superiores só se obtêm de acordo com o Teorema 2.7, pág. 59, se algumas derivadas da função iteradora se anularem no ponto fixo, o que só acontece apenas para funções iteradoras muito particulares.

O método de Newton corresponde precisamente a uma função iteradora particular possuindo a importante vantagem de proporcionar, em geral, convergência de segunda ordem (quadrática). Trata-se de um dos métodos mais frequentemente utilizados, já que combina a rapidez de convergência com a simplicidade do correspondente processo iterativo.

Veremos mais adiante que o método de Newton pode ser encarado, de facto, como um caso particular do método do ponto fixo. Por agora, vamos introduzir este método mediante uma interpretação geométrica.

2.3.1 Interpretação geométrica do método de Newton

Seja f uma função continuamente diferenciável num certo intervalo $[a, b]$. Suponha-se que nesse intervalo a função tem uma única raiz real z e que a sua derivada não se anula (isto é, $f'(x) \neq 0, \forall x \in [a, b]$). Por conseguinte, o ponto z é um *zero simples* da função f.

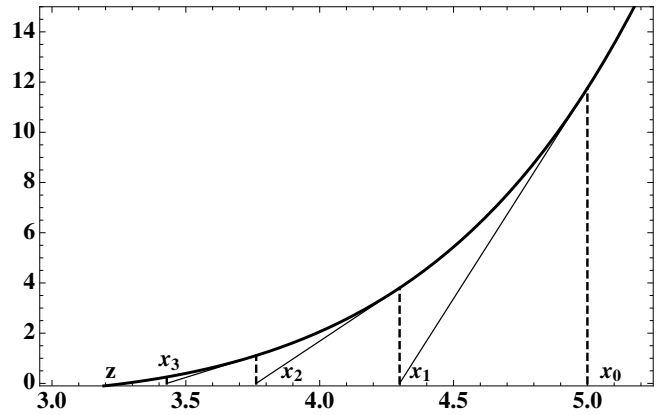


Figura 2.14: Interpretação geométrica do método de Newton.

Sendo x_0 um ponto arbitrário de $[a, b]$, podemos traçar a tangente ao gráfico de f que passa pelo ponto $(x_0, f(x_0))$ (ver Figura 2.14). Sendo $f'(x_0) \neq 0$, essa recta intersecta o eixo das abcissas num certo ponto $(x_1, 0)$. Para determinar x_1, comecemos por escrever a equação da tangente ao gráfico de f em $(x_0, f(x_0))$:

$$y - f(x_0) = f'(x_0)(x - x_0) \ . \tag{2.41}$$

Fazendo $y = 0$ na equação (2.41), obtém-se a abcissa x_1 procurada,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \ .$$

O ponto x_1 será a primeira iterada do método de Newton. As iteradas seguintes serão obtidas do mesmo modo. Mais precisamente, para determinar x_2, traça-se a tangente ao gráfico de f que passa pelo ponto $(x_1, f(x_1))$, e procura-se o ponto onde essa recta intersecta o eixo das abcissas e assim sucessivamente. Deste modo resulta uma sucessão de pontos $(x_k)_{k\geq 0}$, que podem ser calculados pela fórmula de recorrência

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ . \tag{2.42}$$

A interpretação geométrica anterior sugere-nos que a sucessão obtida converge para a raiz z da equação considerada. Nos parágrafos seguintes vamos demonstrar que de facto assim é.

2.3.2 Estimativa do erro do método de Newton

Em primeiro lugar vamos deduzir uma fórmula que nos permite majorar o erro de cada iterada do método de Newton, admitindo que é conhecido um majorante do erro da iterada anterior.

Supomos que a função f satisfaz no intervalo $[a, b]$ as condições já anteriormente referidas (f é continuamente diferenciável em $[a, b]$, e a sua derivada não se anula neste intervalo). Além disso, admitimos que a segunda derivada de f também é contínua neste intervalo. Seja $(x_k)_{k\geq 0}$ a sucessão das iteradas do método (que se consideram pertencentes ao intervalo $[a, b]$).

Se considerarmos a fórmula de Taylor de f, em torno de x_k, obtém-se

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + \frac{(x - x_k)^2}{2}f''(\xi_k), \tag{2.43}$$

onde $\xi_k \in int(x_k, x)$. Fazendo $x = z$ em (2.43), resulta

$$f(z) = f(x_k) + (z - x_k)f'(x_k) + \frac{(z - x_k)^2}{2}f''(\xi_k) = 0, \tag{2.44}$$

com $\xi_k \in int(x_k, z)$. Uma vez que, por hipótese, $f'(x_k) \neq 0$, podemos dividir ambos os membros de (2.44) por $f'(x_k)$, obtendo assim

$$\frac{f(x_k)}{f'(x_k)} + (z - x_k) + \frac{(z - x_k)^2}{2f'(x_k)}f''(\xi_k) = 0 \ . \tag{2.45}$$

Atendendo à fórmula iterativa (2.42) do método de Newton, da equação (2.45) resulta

$$z - x_{k+1} = -\frac{(z - x_k)^2}{2f'(x_k)}f''(\xi_k) \ . \tag{2.46}$$

A igualdade (2.46) fornece a relação que procurávamos entre o erro de x_{k+1} (isto é, e_{k+1}) e o erro de x_k (ou seja, e_k). No segundo membro desta desigualdade aparece o valor $f''(\xi_k)$, o qual não podemos calcular exactamente, já que sabemos apenas que ξ_k é um ponto situado entre x_k e z. Por isso, para podermos majorar o erro absoluto de x_k, ou seja ($|e_k|$), precisamos de majorar o módulo da segunda derivada de f (que se supõe contínua).

Considerando

$$M = \max_{x\in[a,b]} |f''(x)| \tag{2.47}$$

da igualdade (2.46) obtém-se a seguinte relação,

$$|e_{k+1}| \leq |e_k|^2 \frac{M}{2|f'(x_k)|} \ . \tag{2.48}$$

Saliente-se que na desigualdade (2.48) o erro $|e_{k+1}|$ é comparado com o quadrado de $|e_k|$, o que indica um rápido decrescimento do erro. Seja

$$\mu = \min_{x\in[a,b]} |f'(x)| \ . \tag{2.49}$$

A desigualdade (2.48) pode ser reforçada substituindo $|f'(x_k)|$ por μ,

$$|e_{k+1}| \leq |e_k|^2 \frac{M}{2\mu} \ . \tag{2.50}$$

Nesta última desigualdade o segundo membro não depende de k. Na prática, usam-se frequentemente as fórmulas (2.48) e (2.50) para obter uma estimativa de $|e_{k+1}|$.

Exemplo 2.19. *Consideremos a equação*

$$f(x) = \cos(x) - 2\,x = 0,$$

já analisada no Exercício 2.11, pág. 51.
Pretende-se obter aproximações da raiz da equação, situada no intervalo $[0.4, 0.5]$*, mediante aplicação do método de Newton, bem como majorantes do respectivo erro.*

Sendo $x_0 = 0.4$, da fórmula (2.42) obtém-se

$$x_1 = 0.45066547 \quad \text{e} \quad x_2 = 0.45018365 \ .$$

Calculemos majorantes para os erros $|e_1|$ e $|e_2|$. Em primeiro lugar, note-se que $|e_0| \leq 0.5 - 0.4 = 0.1$.

Para podermos aplicar a desigualdade (2.48) é necessário majorar $|f''(x)|$ e minorar $|f'(x)|$. Temos $f'(x) = -\sin(x) - 2$ e $f''(x) = -\cos(x)$. Logo,

$$\mu = \min_{x\in[0.4,0.5]} |f'(x)| = \min_{x\in[0.4,0.5]} |2 + \sin x| = 2 + \sin 0.4 = 2.389,$$

$$M = \max_{x\in[0.4,0.5]} |f''(x)| = \max_{x\in[0.4,0.5]} |\cos x| = \cos 0.4 = 0.921 \ .$$

Por conseguinte, da desigualdade (2.50) resulta a seguinte majoração para o erro absoluto de x_1:

$$|e_1| \leq \frac{M}{2\mu}|e_0|^2 \leq \frac{0.921}{2 \times 2.389} 0.01 = 0.001927.$$

Em relação ao erro de x_2, obtém-se, do mesmo modo,

$$|e_2| \leq \frac{M}{2\mu}|e_1|^2 \leq \frac{0.921}{2 \times 2.389} 0.001927 = 0.696 \times 10^{-7}.$$

Vemos assim que bastam duas iteradas para se conseguir obter um resultado com precisão assaz razoável.

Em complemento apresentamos a seguir uma tabela onde se comparam os resultados obtidos mediante aplicação dos métodos de Newton e do ponto fixo (para

a função iteradora $g(x)\cos(x)/2)$, convidando-se o leitor a verificar os resultados obtidos.

Da análise dos erros que constam da tabela, constata-se imediatamente que o método de Newton possui uma convergência muito mais rápida do que o método de ponto fixo adoptado.

k	x_k (Ponto fixo)	$\lvert e_k\rvert$	x_k (Newton)	$\lvert e_k\rvert$
0	0.4	0.0501	0.4	0.0501
1	0.46053	0.0105	0.45066547	0.48×10^{-3}
2	0.44791	0.0022	0.45018365	0.4×10^{-7}

Comparação entre o método de Newton e o método do ponto fixo (Exemplo 2.19).

Em particular, pode observar-se que para o método de Newton o número de algarismos significativos aproximadamente duplica de uma iteração para a seguinte.

♦

2.3.3 Condições suficientes de convergência

Até ao momento analisámos o erro do método de Newton partindo do princípio de que a aproximação inicial é tal que as iteradas convergem para a raiz procurada. No entanto, nem sempre é fácil prever, para uma dada aproximação inicial, se o método vai ou não convergir e, convergindo, para que raiz se dará tal convergência (caso a equação possua várias raízes).

Neste parágrafo vamos enunciar um conjunto de condições que — uma vez satisfeitas, e no caso da aproximação inicial x_0 pertencer a um certo intervalo — o método converge necessariamente para a raiz da equação que se encontra nesse intervalo.

Teorema 2.9. Seja f uma função real definida no intervalo $I = [a, b]$, verificando as condições:

1. f é contínua em I, e $f(a)f(b) < 0$.
2. $f \in C^1([a, b])$, e $f'(x) \neq 0$ em I.
3. $f \in C^2([a, b])$, sendo $f''(x) \geq 0$ ou $f''(x) \leq 0$ em I.
4. $\dfrac{|f(a)|}{|f'(a)|} < b - a, \quad \text{e} \quad \dfrac{|f(b)|}{|f'(b)|} < b - a$.

Nestas condições, qualquer que seja a aproximação inicial $x_0 \in [a, b]$, o método de Newton converge para a única raiz z de f em I, e a sua convergência é supralinear.

Nalgumas situações tem interesse também a seguinte variante do Teorema 2.9.

Teorema 2.10. Suponhamos que f satisfaz as primeiras três condições do Teorema 2.9. Se a aproximação inicial x_0 for tal que

$$f(x_0)f''(x) \geq 0, \ \forall x \in [a,b],$$

o método de Newton converge para a única raiz z de f em $[a,b]$ e a sucessão das iteradas é *monótona*.

Não iremos fazer a demonstração completa dos dois teoremas anteriores, mas apenas investigar o significado e a razão de ser de cada uma das suas condições.

As primeiras condições, como sabemos pelos Teoremas 2.1 e 2.2, pág. 31, garantem que a função considerada tem um único zero em $[a,b]$. Além disso, a segunda condição é essencial para o método de Newton, pois se ela não se verificar (isto é, se a derivada de f se anular nalgum ponto de $[a,b]$), o método de pode não ser aplicável ou pode convergir lentamente.

Quanto à terceira condição, ela significa que no domínio considerado a segunda derivada de f não muda de sinal ou, por outras palavras, a função não tem pontos de inflexão no intervalo I.

Para entendermos a razão de ser da última condição anteriormente referida, analisemos o seguinte exemplo.

Exemplo 2.20. *Consideremos a função*

$$f(x) = x^3 - x,$$

no intervalo $[-0.5, 05]$. *Poderá garantir convergência do método de Newton para o zero real (único) da função* f, *que existe nesse intervalo?*

No intervalo considerado a função é continuamente diferenciável, com $f'(x) = 3\,x^2 - 1$. Além disso, f possui sinais opostos nos extremos do intervalo ($f(-0.5) = 3/8$, $f(0.5) = -3/8$) e f' não se anula (pois é sempre negativa). Por conseguinte, as duas primeiras condições do Teorema 2.10 estão satisfeitas no intervalo $[-0.5, 0.5]$.

Em relação à terceira condição, temos $f''(x) = 6\,x$, logo $f''(x)$ muda de sinal em $x = 0$, pelo que esta condição não é satisfeita.

Vejamos agora que a convergência do método de Newton não está garantida se tomarmos uma qualquer aproximação inicial no intervalo $[-0.5, 0.5]$.

Seja $x_0 = 1/\sqrt{5} \simeq 0.447214$. Embora este ponto pertença ao intervalo considerado, verifica-se imediatamente que as iteradas do método formam uma sucessão não convergente:

$$\begin{aligned} x_1 &= -1/\sqrt{5} \\ x_2 &= 1/\sqrt{5} \\ x_3 &= -1/\sqrt{5}, \ldots \end{aligned}$$

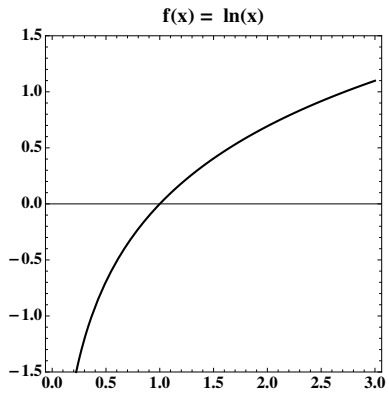


Figura 2.15: Aplicação do método de Newton à equação $\ln(x) = 0$.

⧫

O exemplo a seguir ilustra a importância da hipótese 4 do enunciado do Teorema 2.9.

Exemplo 2.21. *Seja*

$$f(x) = \ln(x).$$

A equação anterior tem uma única raiz $z = 1$ (ver Figura 2.15). Poderá garantir convergência do método para a raiz, independentemente da escolha que fizer da aproximação inicial x_0?

Se considerarmos, por exemplo, o intervalo $[0.5, 3]$, vemos que neste intervalo estão satisfeitas as primeiras 3 condições dos Teoremas 2.9 e 2.10 :

1. $f(0.5) \times f(3) < 0$;
2. $f'(x) = 1/x \neq 0, \quad \forall x \in [0.5, 3]$;
3. $f''(x) = -1/x^2 < 0, \quad \forall x \in [0.5, 3]$.

No entanto, a convergência do método de Newton não está assegurada uma vez escolhida uma qualquer aproximação inicial neste intervalo.

Se tomarmos, por exemplo, $x_0 = 3$, temos $x_1 = 3 - 3\ln(3) < 0$, pelo que o método não pode ser aplicado (visto que $f(x)$ não está definida para $x < 0$).

Neste caso é fácil ver que falha a condição 4 do Teorema 2.9. Com efeito, temos

$$\frac{|f(3)|}{|f'(3)|} = 3\ln(3) > 3 - 0.5 = 2.5 \ .$$

Porém, se escolhermos por exemplo $x_0 = 0.5$, são satisfeitas as condições do Teorema 2.10 (note que $f(0.5) \times f''(x) > 0, \ \forall x \in [0.5, 3]$), pelo que o método de Newton converge para a raiz procurada. ⧫

Sobre o significado geométrico da condição 4 do Teorema 2.9, podemos dizer o seguinte: se ela se verificar, tomando $x_0 = a$, a iterada x_1 satisfaz

$$|x_1 - a| = \frac{|f(a)|}{|f'(a)|} < |b - a|,$$

ou seja, a distância de x_1 a a é menor que o comprimento do intervalo $[a, b]$. Logo, x_1 pertence a esse intervalo. Repetindo este raciocínio pode mostrar-se que todas as iteradas seguintes continuam a pertencer ao intervalo $[a, b]$.

Se começarmos o processo iterativo a partir de $x_0 = b$ e utilizarmos a condição $\frac{|f(b)|}{|f'(b)|} < |b - a|$, um raciocínio semelhante leva-nos à mesma conclusão. Isto é, a condição 4. do Teorema 2.11 garante que *se $x_0 \in [a, b]$, todas as iteradas do método de Newton se mantêm dentro desse intervalo.*

2.3.4 Ordem de convergência do método de Newton

O método de Newton pode ser encarado como um caso particular do método do ponto fixo. Esta abordagem tem a vantagem de permitir analisar a convergência do método de Newton com base nos resultados teóricos que já conhecemos com respeito ao método do ponto fixo.

Consideremos a equação $f(x) = 0$, e suponhamos que existe uma única raiz simples num certo intervalo $[a, b]$. Admitamos ainda que $f \in C^1([a, b])$, e que $f'(x) \neq 0, \forall x \in [a, b]$. A equação considerada é equivalente a

$$x - \frac{f(x)}{f'(x)} = x\,. \tag{2.51}$$

Se definirmos a função iteradora

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

podemos dizer que a equação (2.51) é a equação dos pontos fixos de g. Logo, as raízes de f, que também são pontos fixos de g, podem ser eventualmente aproximadas pelo processo iterativo

$$x_{k+1} = g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}\ . \tag{2.52}$$

Verificamos portanto que este método é idêntico ao método de Newton, aplicado à função $f(x)$. Logo, para determinar a sua ordem de convergência basta determinar, com base no Teorema 2.8, pág. 63, a ordem de convergência da sucessão gerada por esta função iteradora.

Para o efeito, comecemos por calcular as suas derivadas. Temos

$$g'(x) = \frac{f(x)f''(x)}{f'(x)^2} \ .$$

Tomando em consideração que $f(z) = 0$ e $f'(z) \neq 0$, resulta que $g'(z) = 0$. Isto significa que z é ponto fixo *superatractor* para a função iteradora g.
Quanto à segunda derivada de g, temos

$$g''(x) = \frac{(f'(x)f''(x) + f(x)f'''(x))\, f'(x)^2 - f(x)f''(x)(f'(x)^2)'}{f'(x)^4}.$$

Logo,

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

Convergência supralinear

Seja z um zero simples da função f. Do que acima se disse, podemos concluir o seguinte:

a) Se $f''(z) \neq 0$, então $g''(z) \neq 0$ (uma vez que por hipótese $f'(z) \neq 0$). Nesse caso, de acordo com o Teorema 2.8, pág. 63, o método de Newton (ou seja, o método do ponto fixo com a função iteradora $g(x) = x - \dfrac{f(x)}{f'(x)}$) possui ordem de convergência 2 (convergência quadrática). Além disso, a constante assimptótica de convergência é dada por

$$k_\infty = \frac{|f''(z)|}{2\,|f'(z)|} \ .$$

b) Se $f''(z) = 0$, então $g''(z) = 0$, e o método de Newton tem ordem de convergência, pelo menos, 3 (para decidir qual a ordem concreta é necessário analisar as derivadas de ordem superior de g).

Exemplo 2.22. *Considere a equação*

$$f(x) = x^3 - x = 0 \ .$$

Uma das raízes da equação é $z = 0$. Qual é a ordem de convergência do método de Newton aplicado à função em causa, se partir de uma aproximação inicial x_0 suficientemente próxima de z?

Se aplicarmos o método de Newton para o cálculo aproximado desta raiz, tal equivale a utilizar o método do ponto fixo com a função iteradora

$$g(x) = x - \frac{f(x)}{f'(x)} = \frac{2\,x^3}{3\,x^2 - 1} \ .$$

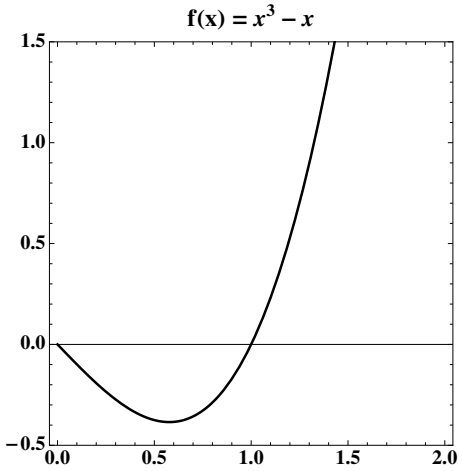


Figura 2.16: Aplicação do método de Newton à equação $x^3 - x = 0$.

Analisemos a ordem do método neste caso. Para isso comecemos por verificar que $f'(0) = -1 \neq 0$ e $f''(0) = 0$. Então, de acordo com a análise que acabámos de realizar, o método deve ter ordem pelo menos 3.
Sabemos que

$$g''(0) = \frac{f''(0)}{f'(0)} = 0 \ .$$

Para determinar $g'''(0)$, observemos que a função g admite, em torno de $z = 0$, um desenvolvimento de Taylor da forma

$$g(x) = \frac{-2\,x^3}{1 - 3\,x^2} = -2\,x^3 + O(x^5),$$

de onde se conclui que $g'''(x) = -12 + O(x^2)$, pelo que $g'''(0) = -12$. Temos, portanto, convergência de ordem 3.

A constante assimptótica de convergência, de acordo com o Teorema 2.7, é

$$k_\infty = \frac{|g'''(0)|}{3!} = 2 \ .$$

Inspecccionando o gráfico da função iteradora de Newton

$$g(x) = x - f(x)/f'(x) = \frac{2\,x^3}{3\,x^2 - 1},$$

(ver Figura 2.17), facilmente se reconhece que o método de Newton, uma vez escolhido um ponto inicial x_0 próximo de cada um dos pontos fixos de g, a rapidez de convergência do método será maior num caso do que noutro. Porquê?

Sugere-se ao leitor que experimente o que acontece se usar a função iteradora de Newton, partindo de $x_0 \simeq \pm 1/\sqrt{3} \simeq \pm 0.58$. ♦

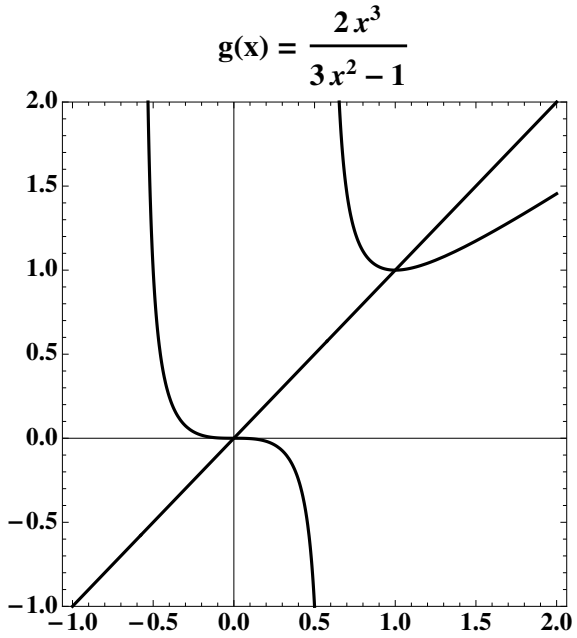


Figura 2.17: Função iteradora para o Exemplo 2.22.

2.3.5 Método de Wegstein *

No Exemplo 2.23 a seguir é definida uma nova função iteradora que define um processo iterativo designado por método de Wegstein.

Exemplo 2.23.

A fim de aproximar o número π, considere a equação

$$f(x) = tan(x/4) - 1 = 0, \quad x \in I = [2, 4] \ .$$

(a) Seja g a função iteradora de Newton aplicada a f. Mostre que são válidas as condições de aplicabilidade do teorema do ponto fixo para g, no intervalo I. Diga, justificando, qual a ordem de convergência do método de Newton quando inicia o processo em $x_0 = 2$.

(b) Desenvolva um ou mais programas a fim de produzir uma tabela como a dada na Figura 2.18. Os resultados numéricos que obteve sugerem concordância com a alínea anterior? Justifique.
Pretende-se comparar o método anterior com outro designado por *método de Wegstein*[1][8]. Este método é definido através de uma função iteradora W, cujas propriedades geométricas são ilustradas na Figura 2.19.

[8](1) J. H. Wegstein, Accelerating convergence of iterative processes, *Commun. ACM* 1, 9-13 (1958).

Newton	Erro
2.00000000000	1.14159265359
3.39766264212	-0.25606998853
3.15866228243	-0.01706962884
3.14166570344	-0.00007304985
3.14159265492	-1.33×10^{-9}

Figura 2.18: 4 iterações do método de Newton iniciado com $x_0 = 2$.

A função W, designada por iteradora de Wegstein, depende da função iteradora de partida g. Tal como mostrado na referida figura, a iteradora de Wegstein transforma o ponto x no ponto $W(x)$. O declive do segmento de recta que une os pontos $A = (x, g(x))$ e $B = (g(x), g^2(x))$ figurados é

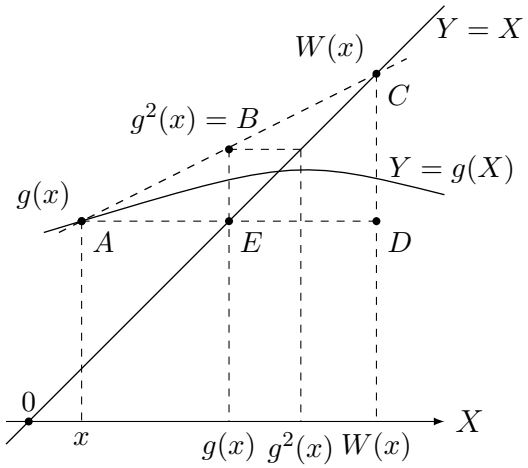


Figura 2.19: Função iteradora do método de Wegstein.

$$m(x) = \frac{g^2(x) - g(x)}{g(x) - x}, \quad \text{onde} \quad g(x) \neq x, \quad \text{e} \quad g^2(x) = g(g(x)) \ .$$

O ponto C está localizado sobre a recta de equação $Y = X$, e tem para coordenadas $(W(x), W(x))$. Assim, atendendo à semelhança dos triângulos $\triangle ACD$ e $\triangle ABE$, pode deduzir-se que

$$m(x) = \frac{W(x) - g(x)}{W(x) - x} = \frac{g^2(x) - g(x)}{g(x) - x}, \quad (*)$$

$$\text{onde} \quad W(x) \neq x \ .$$

Da primeira igualdade em $(*)$ resulta

$$(m(x) - 1)\, W(x) = m(x)\, x - g(x) \ .$$

Admitindo que $m(x) \neq 1$, da igualdade anterior obtém-se

$$W(x) = \frac{1}{1-m(x)}\, g(x) + \frac{m(x)}{m(x)-1}\, x, \quad \text{onde} \quad m(x) = \frac{g^2(x) - g(x)}{g(x) - x} \ . \quad (2.53)$$

(c) Usando a definição da função iteradora de Wegstein dada em (2.53) e tomando para g a função iteradora de Newton (relativamente à função f considerada anteriormente), desenvolva um ou mais programas a fim de produzir uma tabela análoga à da Figura 2.20.

Os resultados numéricos sugerem que o número de algarismos significativos de cada iterada do método de Wegstein aproximadamente triplica de iteração para iteração. Poderá fazer alguma conjectura a respeito da ordem de convergência deste método quando aplicado ao problema em causa? Justifique.

Wegstein	Erro
2.000000000000000000000000000000000000000000000000000000000000	1.141592654
3.193563280297456657811812531835980456300589497064541692712 83	-0.05197062671
3.141583607742694116347410754312766549742639840852167144520 53	$9.045847099 \times 10^{-6}$
3.141592653589793284724755162260968741001946350425863204708 95	$-4.626211178 \times 10^{-17}$
3.141592653589793238462643383279502884197169399375099632888 41	$6.18808653 \times 10^{-51}$

Figura 2.20: 4 iterações do método de Wegstein iniciado com $x_0 = 2$.

(d) Aplicando rotinas do seu sistema computacional escreva um programa que lhe permita obter uma ilustração das funções iteradoras de Newton e de Wegstein análoga à mostrada na Figura 2.21.

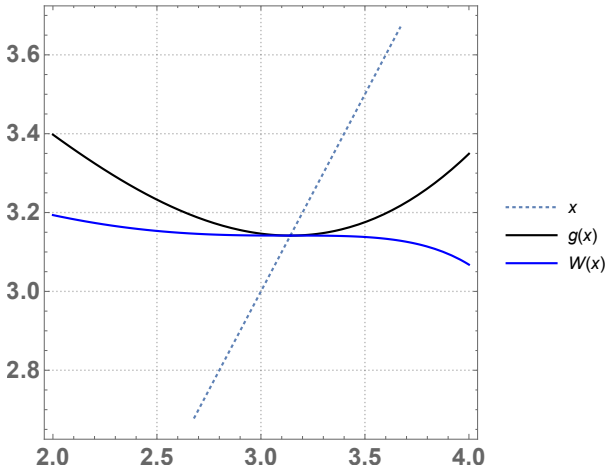


Figura 2.21: Funções iteradoras de Newton e de Wegstein para $f(x) = tan(x/4) - 1, \quad x \in [2, 4]$.

(e) Diga, justificando, se da observação da Figura 2.21 é ou não previsível que, para $x_0 = 2$, a partir da primeira iterada do método de Newton, o erro mantenha sinal (cf. tabela da Figura 2.18), enquanto que para o método de Wegstein o respectivo erro é oscilante (cf. tabela da Figura 2.20).
Se quisesse demonstrar que o método de Wegstein, aplicado ao problema em causa, é de terceira ordem de convergência, que resultado(s) teórico(s) iria aplicar?

(f) Dada uma função iteradora g, tal que z é ponto fixo de g e $g'(z) \neq 1$, seja $\lim_{x->z} m'(x) = \alpha$. Sabe-se que, se o método de Wegstein for convergente para z e

$$\alpha = 1/2\, g''(z), \qquad (**)$$

a sua ordem de convergência é, pelo menos, 3. Para o problema em causa e usando computação simbólica mostre que é satisfeita a igualdade $(**)$.

2.4 Transformação de ponto fixo em superatractor *

No problema da catenária, pág. 29, foi estabelecida uma equação do tipo $\phi(h) = 0$, a partir da qual foi gerado um método do ponto fixo discutido no Exemplo 2.14, pág. 56. Nesse exemplo invoca-se o gráfico de uma certa função iteradora g, para se concluir que o ponto fixo z respectivo é *atractor*, isto é, $|g'(z)| < 1$. Acontece que próximo do ponto fixo $|g'| \simeq 1$, o que deixa prever que o método convergirá lentamente.

Coloca-se a questão de saber se não será possível transformar a função g numa outra função iteradora, de modo que z seja ponto fixo *superatractor* para esta nova função. Vejamos como esse objectivo poderá ser realizado usando devidamente o método de Newton. Supomos que todas as funções envolvidas são suficientemente regulares numa vizinhança do ponto fixo.

Com efeito, a partir de uma certa função iteradora h, seja z um seu ponto fixo tal que $|h'(z)| > 1$ (ou $|h'(z)| < 1$ mas $|h'(z)| \simeq 1$). Considerem-se as funções f e g, assim definidas:

$$f(x) = h(x) - x$$

e

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{h(x) - x}{h'(x) - 1}\ . \qquad (2.54)$$

Uma vez que por hipótese $|h'(z)| \neq 1$, o denominador na fracção que se encontra no segundo membro da igualdade anterior é não nulo para x próximo de z. Atendendo a que

$$g'(x) \ = 1 - \frac{(h'(x) - 1)^2 - (h(x) - x)\, h''(x)}{(h'(x) - 1)^2} = \frac{(h(x) - x)\, h''(x)}{(h'(x) - 1)^2},$$

e a que z é ponto fixo de h, obtém-se

$$g'(z) = \frac{(h(z) - z)\, h''(z)}{(h'(z) - 1)^2} = 0 \ .$$

Assim, caso z seja ponto fixo repulsor para h, o mesmo ponto fixo passa a ser superatractor para g. Note que a função g foi construída aplicando a função iteradora de Newton à funçao $f(x) = h(x) - x$ (evidentemente que a função $f(x) = x - h(x)$ também serve para o efeito).

Exemplo 2.24. *Levando em consideração os dados do exemplo da catenária, pág. 56, definimos a seguinte função iteradora, a qual sabemos possuir um único ponto fixo no intervalo* $[0, 50]$*,*

$$\phi(h) = (10 + h) \cosh\left(\frac{20}{10 + h}\right) - 15 \ .$$

Pretende-se transformar ϕ numa outra função iteradora de convergência mais rápida.

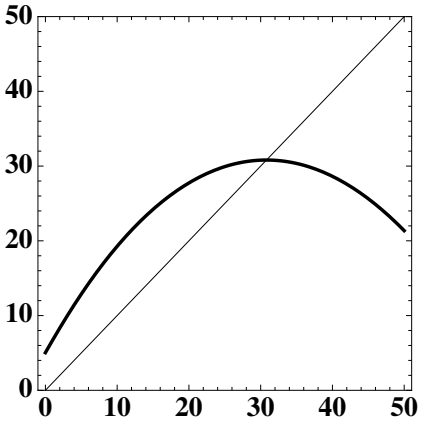
Consideremos a função $f(h) = h - \phi(h)$. Transformando esta função na respectiva função iteradora de Newton $g_1(h) = h - f(h)/f'(h)$, resulta

$$g_1(h) = h - \frac{h - (h + 10) \cosh\left(\dfrac{20}{h + 10}\right) + 15}{\dfrac{20 \sinh\left(\dfrac{20}{h + 10}\right)}{h + 10} - \cosh\left(\dfrac{20}{h + 10}\right) + 1} \ .$$

Por exemplo, fixada a aproximação inicial $h_0 = 10$, encontra na Figura 2.22 o gráfico da função iteradora g_1 acompanhado por uma tabela de iteradas dessa função, começando com h_0. Note que o ponto fixo é superatractor para a função g_1.

Podemos portanto concluir que a altura h pretendida é de $30\,807\, mm$. Dado que o parâmetro da catenária vale aproximadamente $a \simeq 40.8\, m$ (valor obtido pelo método da bissecção, pág. 40), uma vez que $a = d + h$ e $d = 10\, m$, resulta $h = a - d \simeq 30.8\, m$, uma aproximação que é consistente com as aproximações calculadas na tabela da Figura 2.22.

Note que na referida tabela as iteradas aparecem com um número decrescente de algarismos significativos. Esse decrescimento fica a dever-se ao facto da expressão dada para a função iteradora g_1 estar sujeita ao efeito de cancelamento subtractivo, à medida que h se aproxima do ponto fixo. No entanto, tem em vista a precisão de h requerida, a resposta ao problema inicialmente proposto pode ser dada através do último valor tabelado. ♦



10.00000000000000
19.2729347502438
27.316518871489
30.49025366785
30.8045283340
30.807132511
30.8071327

Figura 2.22: Função iteradora transformada de uma função iteradora lenta.

2.5 Método da secante

Tal como no caso do método de Newton, a fórmula iterativa deste método vai ser deduzida a partir de uma interpretação geométrica.

2.5.1 Interpretação geométrica do método da secante

Seja f uma função real, contínua num certo intervalo $[a, b]$, e suponha-se que f tem nesse intervalo um único zero z. Para aplicar o método da secante, escolhem-se dois números, x_0 e x_1, no intervalo $[a, b]$, e considera-se a recta que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$ (secante ao gráfico de f). A equação dessa recta é

$$y - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \ .$$

Depois, determina-se o ponto onde esta recta intersecta o eixo das abcissas. A intersecção desta recta com o eixo das abcissas existe desde que $f(x_0) \neq f(x_1)$, condição que consideramos satisfeita. Designando por x_2 a abcissa desse ponto, obtém-se a seguinte equação para x_2,

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) \ . \tag{2.55}$$

Considera-se x_2 como sendo a nova aproximação da raiz, definida a partir de x_0 e x_1.

A fórmula que nos permite determinar cada aproximação x_{k+1}, a partir das duas anteriores x_k e x_{k-1}, é análoga a (2.55),

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \qquad k = 1, 2, \ldots \tag{2.56}$$

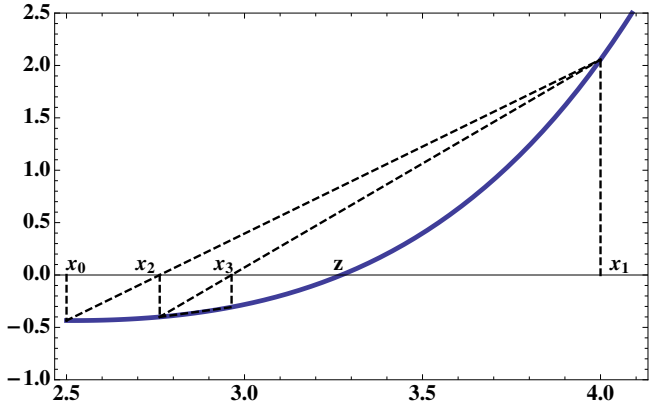


Figura 2.23: Interpretação geométrica do método da secante.

Uma ilustração do método da secante está representada na Figura 2.23.

2.5.2 Estimativa de erro

No caso do método de Newton, vimos que o erro de cada iterada pode ser estimado a partir do erro da iterada anterior e das propriedades da função f. Para o método da secante é de realçar uma diferença fundamental: cada iterada depende das duas iteradas anteriores e não apenas da última. Neste caso, diz-se que temos um *método iterativo a dois passos*.

Sendo assim, é natural que o erro de cada iterada do método da secante possa ser determinado a partir dos erros das duas últimas iteradas.

Suponhamos então que x_{m-1} e x_m são duas iteradas consecutivas do método. A iterada seguinte, x_{m+1}, é determinada através da fórmula (2.56). Representemos os erros de x_{m-1} e x_m respectivamente por e_{m-1} e e_m, isto é, $e_{m-1} = z - x_{m-1}$ e $e_m = z - x_m$. Além disso, suponhamos que a função f é duas vezes continuamente diferenciável num intervalo $I \subset [a, b]$, que contém x_{m-1}, x_m, x_{m+1} e z, e que f' não se anula em I.

Pode mostrar-se ([1], pág. 67), que e_{m+1} (erro de x_{m+1}) satisfaz a desigualdade,

$$e_{m+1} = -\frac{f''(\xi_m)}{2f'(\eta_m)} e_m e_{m-1}, \tag{2.57}$$

onde ξ_m e η_m representam pontos que pertencem ao intervalo I acima referido.

Note-se que a fórmula (2.57) é semelhante à fórmula (2.46) para o erro do método de Newton, da pág. 67. A diferença consiste, como seria de esperar, ser o erro da nova iterada do método da secante avaliado a partir do produto dos erros das duas últimas iteradas, enquanto que no método de Newton o erro da nova iterada é avaliado a partir do quadrado do erro da iterada anterior.

Majorações de erro

À semelhança do que fizemos no caso do método de Newton, para usar a fórmula (2.57) convém majorar (no intervalo I) o módulo da segunda derivada de f e minorar o módulo da sua primeira derivada. Para simplificar, suponhamos que $I = [a, b]$, e

$$M = \max_{x\in[a,b]} |f''(x)|, \quad \text{e} \quad \mu = \min_{x\in[a,b]} |f'(x)| \ .$$

Da fórmula (2.57) resulta imediatamente a seguinte majoração para o erro absoluto do método da secante,

$$|z - x_{m+1}| = |e_{m+1}| \leq \frac{M}{2\,\mu}\,|e_m|\,|e_{m-1}| \ . \tag{2.58}$$

Normalmente, os erros absolutos das duas iteradas iniciais, $|e_0|$ e $|e_1|$, são majorados pelo comprimento do intervalo $[a, b]$. Isto é, são evidentes as desigualdades $|e_0| < |b - a|$ e $|e_1| < |b - a|$. A partir daí os erros das sucessivas iteradas são majorados por recorrência, isto é, o erro $|e_2|$ majora-se a partir dos erros $|e_0|$ e $|e_1|$; o erro $|e_3|$ majora-se a partir dos erros $|e_1|$ e $|e_2|$; e assim sucessivamente.

Estimativas e majorações de erro para as sucessivas aproximações calculadas pelo método da secante podem obter-se mais facilmente aplicando as fórmulas (2.5) e (2.6), pág. 34, conforme ilustrado no Exemplo (2.25).

Exemplo 2.25. *Consideremos mais uma vez a equação*

$$f(x) = \cos(x) - 2\,x = 0,$$

a qual possui uma raiz no intervalo $[0.4, 0.5]$. *Para aproximar essa raiz pretende-se usar o método da secante.*

(a) Tomando como aproximações iniciais os pontos $x_0 = 0.5$ *e* $x_1 = 0.4$, *calculemos as iteradas* x_2 *e* x_3 *pelo método da secante.*

(b) Determinem-se majorantes do erro absoluto de x_0, x_1, x_2 *e* x_3 *e estimativas, afectadas de sinal, desses erros.*

(a) Aplicando a fórmula (2.56), temos

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) = 0.449721$$

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 0.450188 \ .$$

(b) O caminho mais fácil seria majorar $|e_0|$ e $|e_1|$ considerando o comprimento do intervalo em causa, $|b - a| = |0.5 - 0.4| = 0.1$. O majorante pode, no entanto, ser

um pouco melhorado se tivermos em conta o sinal de f em cada um dos pontos x_i calculados. Para tal, observemos a tabela

i	x_i	$f(x_i)$
0	0.5	-0.122
1	0.4	0.121
2	0.449721	0.0011
3	0.450188	-0.00001

Da tabela anterior conclui-se que os pontos x_1 e x_2 se encontram à esquerda da raiz z (onde f é positiva), enquanto x_0 e x_3 se encontram à direita (onde f é negativa). Sendo assim, para os erros de x_0 e x_1 obtêm-se os seguintes majorantes:

$$|e_0| = |z - x_0| \leq |x_2 - x_0| = |0.449721 - 0.5| = 0.050258,$$

$$|e_1| = |z - x_1| \leq |x_3 - x_1| = |0.450188 - 0.4| = 0.050188 \ .$$

Recordemos do Exemplo 2.19, pág. 68, que neste caso se tem $M = 0.921$, $\mu = 2.389$. Assim, pela estimativa (2.58), obtém-se

$$|e_2| \leq \frac{M}{2\mu}|e_1||e_0| \leq 0.193 \times 0.050188 \times 0.050258 = 0.4868 \times 10^{-3},$$

$$|e_3| \leq \frac{M}{2\mu}|e_2||e_1| \leq 0.193 \times 0.4868 \times 10^{-3} \times 0.050188 = 0.4715 \times 10^{-5}.$$

Vemos assim que, ao fim de duas iterações, o método da secante nos proporciona uma aproximação com um erro da ordem de 10^{-5}. No caso de método de Newton, com o mesmo número de iterações, obtém-se um erro da ordem de 10^{-7} (ver Exemplo 2.19, pág. 68).

Apliquemos agora as fórmulas (2.5) e (2.6), pág. 34 .

Como

$$m = \min_{0.4 \leq x \leq 0.5} |f'(x)| = 2 - \sin(0.4) > 0,$$

resultam respectivamente as estimativas e majorações de erro

$$\bar{e}_k = -\frac{f(x_k)}{f'(x_k)} = \frac{\cos(x_k) - 2\,x_k}{2 + \sin(x_k)}, \quad k = 0, 1, \ldots,$$

e

$$|e_k| = |z - x_k| \leq \frac{|f(x_k)|}{m} = \frac{|\cos(x_k) - 2\,x_k|}{2 - \sin(0.4)}, \quad k = 0, 1, \ldots \ .$$

O resultado da aplicação das duas fórmulas anteriores para as aproximações x_0 a x_3 é mostrado na tabela a seguir. Os valores tabelados confirmam a ordem de

grandeza das majorações de erro calculadas através de (2.58).

k	x_k	$\bar{e}_k = -\dfrac{f(x_k)}{f'(x_k)}$	$\lvert e_k\rvert \leq \dfrac{\lvert f(x_k)\rvert}{m}$
0	0.5	-0.0493733	0.0512332
1	0.4	0.0506650	0.0506655
2	0.449721	0.0004627	0.0004714
3	0.450188	-4.4×10^{-6}	4.5×10^{-6}

Atendendo a que $z = 0.45018361129\cdots$, pode verificar-se que os valores da terceira coluna da tabela são, de facto, boas estimativas de erro. ♦

O exemplo anterior sugere que o método de Newton converge mais rapidamente do que o da secante. Por outro lado, já vimos anteriormente que a precisão que se consegue obter com duas iteradas do método do ponto fixo é da ordem de 10^{-2}. Estas observações sugerem ser de esperar que a ordem de convergência do método da secante esteja entre a ordem do método do ponto fixo (usualmente de ordem um de convergência) e a do método de Newton (usualmente de ordem dois). Esta conjectura é confirmada pelo estudo que efectuamos de seguida.

2.5.3 Convergência do método da secante

Com base na estimativa do erro que foi deduzida no parágrafo anterior, pode provar-se o seguinte teorema sobre a convergência do método da secante (ver demonstração em [1], pág. 69).

Teorema 2.11. Seja f uma função duas vezes continuamente diferenciável numa vizinhança de z, tal que $f'(z) \neq 0$. Se os valores iniciais x_0 e x_1 forem suficientemente próximos de z, a sucessão $(x_m)_{m\geq 0}$ gerada pelo método da secante converge para z.

Como se disse ao discutir o Exemplo 2.25, o método da secante aparenta ser mais rápido que o método do ponto fixo (o qual geralmente tem ordem um), mas menos rápido que o de Newton (que em geral possui convergência quadrática). Com efeito, sob certas condições sobre a função em causa, se (x_m) for uma sucessão gerada pelo método da secante, existe um número real p, tal que $1 < p < 2$, para o qual se verifica

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{|z - x_m|^p} = K_\infty, \tag{2.59}$$

onde K_∞ é uma constante positiva, que de acordo com a Definição (2.4), pág. 60, designa o coeficiente assimptótico de convergência.
Mais precisamente, pode provar-se (ver detalhes em [1]), que

$$p = \frac{1+\sqrt{5}}{2} \approx 1.618,$$

isto é, a ordem de convergência deste método é dada pelo chamado *número de ouro* (sobre a importância desse número e as suas numerosas aplicações ver, por exemplo, [20]).

O Teorema 2.11 anterior tem a desvantagem de não ser facilmente aplicável. Na realidade, o que significa a frase "se x_0 e x_1 forem suficientemente próximos de z"?

Na prática são bastante mais úteis resultados como os anunciados a seguir, os quais são do tipo dos Teoremas 2.9 e 2.10, pág. 70. Estes proporcionam condições suficientes para a convergência do método da secante, desde que as aproximações iniciais pertençam a um dado intervalo. Passamos a enunciar esses teoremas.

Teorema 2.12. Nas condições do Teorema 2.9, pág. 69, o método da secante converge para a raiz z de f em $[a, b]$, quaisquer que sejam as aproximações iniciais x_0,x_1, pertencentes a $[a, b]$.

Teorema 2.13. Se as primeiras três condições do Teorema 2.9 se verificam e as aproximações iniciais satisfizerem

$$f(x_0)\, f''(x) \geq 0 \quad \text{e} \quad f(x_1)\, f''(x) \geq 0, \ \forall x \in [a, b],$$

então o método da secante converge para a raiz z de f em $[a, b]$.

2.5.4 Estimativas realistas de erro *

Supondo que um determinado processo iterativo converge para uma raiz z, fórmulas de majoração de erro como (2.50), pág. 68, possuem o inconveniente de exigirem um esforço analítico e de cálculo, em geral superiores ao esforço exigido pelo método iterativo propriamente dito. Acresce que essas majorações de erro com frequência sobrestimam o erro realmente cometido.

Por conseguinte, interessa dispor de um processo expedito para obter *estimativas* do erro de uma iterada, $e_k = z - x_k$, utilizando se possível um ou mais valores de iteradas já calculadas, de modo a obter-se uma estimativa *realista* do erro e_k. Neste sentido, o teorema de Lagrange é de grande utilidade, conforme foi discutido no parágrafo 2.1.2, pág. 33, que relembramos a seguir.

Admitindo que z é uma raiz simples da equação $f(x) = 0$, onde f é suficientemente regular numa vizinhança de z, e que x_k é uma aproximação de z calculada mediante um certo processo iterativo, pelo teorema de Lagrange, temos

$$f(z) = f(x_k) + f'(\xi_k)\,(z - x_k), \qquad \xi_k \in \ int(x_k, z)\ .$$

Atendendo a que $f(z) = 0$, tem-se

$$e_k = x - x_k = -\frac{f(x_k)}{f'(\xi_k)}, \quad \xi_k \in \ int(x_k, z)\ . \tag{2.60}$$

	x_k	$z - x_k$	Estimativa (2.62)
x_0	0.4	0.0501836	0.0506655
x_1	0.450665	-0.000481855	-0.000481812
x_2	0.450184	-4.29096×10^{-8}	-4.29096×10^{-8}
x_3	0.450184	-3.33067×10^{-16}	

Tabela 2.1: Estimativas realistas de erro para o método de Newton.

	x_k	$z - x_k$	Estimativa (2.64)
x_0	0.4	0.0501836	
x_1	0.46053	-0.0103469	-0.011040
x_2	0.447908	0.00227518	0.002270
x_3	0.450677	-0.0004938	

Tabela 2.2: Estimativas realistas de erro para o método de ponto fixo.

Como por hipótese f' é função contínua numa vizinhança de z, sendo x_k "próximo"de z, então $f'(\xi_k) \simeq f'(x_k)$, pelo que de (2.60) resulta,

$$e_k = x - x_k \simeq -\frac{f(x_k)}{f'(x_k)} \ . \tag{2.61}$$

A fórmula anterior permite-nos, por exemplo, obter estimativas realistas do erro no método da bissecção, e essa estimativa será tanto mais realista quanto mais próximo a aproximação x_k estiver da raiz z.

A expressão (2.61) encontra aplicação imediata no próprio método de Newton. Com efeito, uma vez que para este método é válida a fórmula recursiva $x_{k+1} = x_k - f(x_k)/f'(x_k)$, comparando com (2.61), resulta

$$\begin{aligned} e_k \ &= z - x_k \simeq x_{k+1} - x_k \\ &\text{(estimativa de erro para método de Newton).} \end{aligned} \tag{2.62}$$

A fórmula aproximada (2.62) diz-nos que é possível calcular uma estimativa realista do erro de uma iterada x_k do método de Newton, usando apenas a informação contida na dupla x_k, x_{k+1}.

Num método do ponto fixo geral, com função iteradora g suficientemente regular numa vizinhança de um ponto fixo z, e tal que $g'(z) \neq 1$ (ou seja, desde que o ponto fixo não seja neutro), vejamos que podemos obter estimativas realistas do erro de uma iterada x_k, à custa da informação contida na tripla x_{k-1}, x_k, x_{k+1}.

Atendendo a que para

$$f(x) = x - g(x),$$

se tem

$$-\frac{f(x)}{f'(x)} = \frac{g(x)-x}{1-g'(x)},$$

a expressão (2.61) pode ser rescrita como

$$e_k = z - x_k \simeq \frac{x_{k+1}-x_k}{1-g'(x_k)} \ . \tag{2.63}$$

Ora, pelo teorema de Lagrange,

$$g(x_k) = g(x_{k-1}) + g'(\xi_{k-1})\,(x_k - x_{k-1}), \quad \xi_{k-1}\ int(x_{k-1}, x_k) \ .$$

Admitindo que g' é contínua numa vizinhança de z, e sendo x_{k-1} e x_k valores próximos de z, tem-se $g'(\xi_{k-1}) \simeq g'(x_k)$. Assim, a expressão (2.63) pode ser substituída pela estimativa de erro

$$e_k = z - x_k \simeq \frac{x_{k+1}-x_k}{1-\dfrac{x_{k+1}-x_k}{x_k-x_{k-1}}} \ . \tag{2.64}$$

Exemplo 2.26. *Voltando ao Exemplo 2.19, pág. 68, seja*

$$f(x) = \cos(x) - 2\,x \Leftrightarrow x = g(x) = \frac{\cos(x)}{2}, \quad \textit{com} \quad z = 0.45018361129487357 \ .$$

Usando como aproximação inicial $x_0 = 0.4$, efectuar três iterações, respectivamente pelo método de Newton aplicado à função f, e pelo método de ponto fixo com função iteradora g. Comparar os respectivos erros exactos com os erros estimados segundo (2.62) *e* (2.64).

As respectivas estimativas realistas de erro são dadas nas tabelas 2.1 e 2.2. ♦

2.6 Exercícios resolvidos

No exercício a seguir é dada uma família de processos iterativos de ponto fixo cuja ordem de convergência é tão grande quanto se queira. Os métodos numéricos subjacentes são úteis para aproximar com alta precisão números da forma $1/\alpha$, sem efectuar divisões.

Exercício 2.1. *Dado o número real positivo $\alpha \neq 1$, pretende-se aproximar o número $z = \dfrac{1}{\alpha}$, mediante um algoritmo sem intervenção da operação de divisão. Para o efeito, considere a família de processos iterativos gerados pelas funções iteradoras $g_1, g_2, g_3, \ldots$, assim definidas:*

$$\begin{aligned} g_1(x) &= x + x\,(1-\alpha\,x) \\ g_2(x) &= x + x\,(1-\alpha\,x) + x\,(1-\alpha\,x)^2 \\ &\vdots \\ g_k(x) &= g_{k-1}(x) + x\,(1-\alpha\,x)^k, \qquad k \geq 2 \ . \end{aligned} \tag{2.65}$$

k	g_k'	$g_k^{(2)}$
1	$-2\,(\alpha\,x-1)$	$-2\,\alpha$
2	$3\,(\alpha\,x-1)^2$	$6\,\alpha\,(\alpha\,x-1)$
3	$-4\,(\alpha\,x-1)^3$	$-12\,\alpha\,(\alpha\,x-1)^2$
4	$5\,(\alpha\,x-1)^4$	$20\,\alpha\,(\alpha\,x-1)^3$
5	$-6\,(\alpha\,x-1)^5$	$-30\,\alpha\,(\alpha\,x-1)^4$
6	$7\,(\alpha\,x-1)^6$	$42\,\alpha,(\alpha\,x-1)^5$
7	$-8\,(\alpha\,x-1)^7$	$-56\,\alpha\,(\alpha\,x-1)^6$

Tabela 2.3: Primeira e segunda derivadas das funções iteradoras (2.65).

Diga, justificando, se são verdadeiras ou falsas as seguintes afirmações (a)–(c):

(a) Para qualquer inteiro $k \geq 1$, os pontos 0 e $1/\alpha$ são pontos fixos da função iteradora g_k.

(b) Se $k = 1$, o ponto fixo $z = 1/\alpha$ é atractor. Leve em consideração a informação contida na Tabela 2.3.

(c) Para $k \geq 2$, o processo iterativo gerado pela função g_k possui ordem de convergência k.

(d) Para $\alpha = \pi$, desenhe os gráficos das funções iteradoras g_k, para $1 \leq k \leq 7$, no intervalo $[0, 1]$.
Escolhido um valor inicial suficientemente próximo do ponto fixo $1/\pi$, por que razão podemos antecipar que a sucessão gerada por g_7 converge muito mais rapidamente para $1/\pi$ do que a sucessão gerada por g_1?

(e) Considere $\alpha = \pi$. Fazendo $x_0 = 1/10$, e usando precisão adequada nos cálculos, aplique a função iteradora g_7 de modo a obter uma aproximação de $z = 1/\pi$, com pelo menos 500 algarismos significativos.

(a) Os pontos fixos da função iteradora g_1 são solução da equação $g_1(x) = x$. Ou seja,

$$x + x\,(1 - \alpha\,x) = x \iff x\,(1 - \alpha\,x) = 0 \iff x = 0 \quad \vee \quad x = 1/\alpha\ .$$

Atendendo às expressões (2.65), para qualquer inteiro $k \geq 1$, os pontos fixos de g_k são solução da equação

$$g_{k-1}(x) + x\,(1 - \alpha\,x)^k = x\ .$$

Assim, se z é ponto fixo de g_{k-1}, resulta da equação anterior

$$z + z\,(1 - \alpha\,z)^k = z \implies z = 0 \quad \vee \quad z = 1/\alpha\ .$$

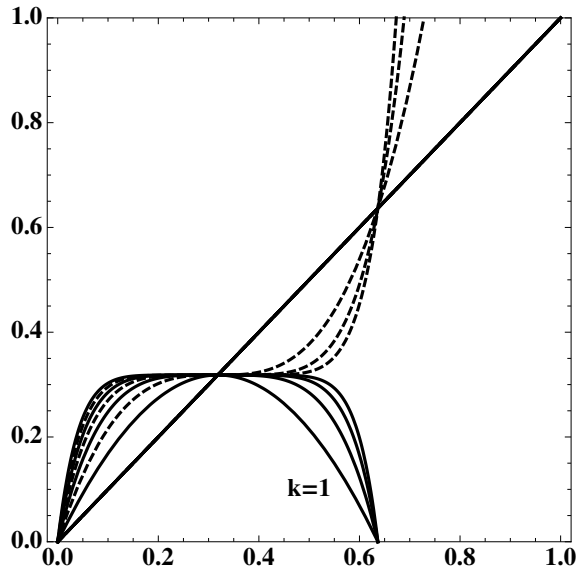


Figura 2.24: Funçoes iteradoras (2.65), com $1 \leq k \leq 7$ e $\alpha = \pi$. A traço cheio para k ímpar e a tracejado para k par.

Como 0 e $1/\alpha$ são pontos fixos da função g_1, conclui-se que esses pontos fixos são também pontos fixos de g_k, para $k \geq 2$.

(b) Dado que $g_1(1/\alpha) = 1/\alpha$, $g_1'(1/\alpha) = 0$ e $g_1''(1/\alpha) = -2\,\alpha \neq 0$, ou seja, o ponto fixo $1/\alpha$ é superatractor para g_1. Escolhido x_0 suficientemente próximo do ponto fixo, o processo $x_{k+1} = g_1(x_k)$ converge para $1/\alpha$. A convergência é de ordem $p = 2$ (ver Teorema 2.8, pág. 63), e o coeficiente assimptótico de convergência é

$$k_\infty = \frac{|g_1''(1/\alpha)|}{2} = \alpha \ .$$

(c) A partir da informação contida na tabela 2.3, conclui-se que para $2 \leq k \leq 7$, são válidas as igualdades

$$\begin{array}{ll} g_k^{(j)}(1/\alpha) = 0, & \text{para} \quad 1 \leq j \leq k-1 \\ g_k^{(k)}(1/\alpha) = (-1)^k\, k!\, \alpha^k \neq 0 \ . & \end{array}$$

Por conseguinte, o processo iterativo respectivo é de ordem k e o coeficiente assimptótico de convergência é

$$k_\infty = \frac{|g_k^{(k)}(1/\alpha)|}{k!} = \alpha^k.$$

Sugere-se ao leitor que use indução matemática para mostrar que o resultado anterior é válido para qualquer número natural k, ou seja, que é arbitrária a ordem de convergência do processo iterativo gerado pela função g_k.

```
0.31830988618379067153776752674502872406891929148091289749533468811779359↖
 52684530701802276055325061719121456854535159160737858236922291573057559 3↖
 48214633996784584799338748181551461554927938506153774347857924347953233 8↖
 67247804834472580236647602284453995114318809237801738053479122409788218 7↖
 38756881710574461998928868004973446954789192217966461935661498123339729 2↖
 56093988973043757631495731339284820779917482786972199677361983999248857 5↖
 11703423577168622350375343210930950739760194789207295186675361186050
```

Figura 2.25: Aproximação de $1/\pi$ com 500 algarismos significativos.

(d) Os gráficos de g_k desenhados na Figura 2.24 mostram que 0 e $1/\alpha \simeq 0.32$ são pontos fixos comuns à funções g_k, para $1 \leq k \leq 7$.

No intervalo considerado, e para $k = 2, 4$ e 6, as respectivas funções iteradoras intersectam a recta $y = x$ num ponto fixo (repulsor) que é distinto dos anteriores. Um tal ponto fixo recebe a designação de *ponto fixo estranho* (assim designado por não ser ponto fixo da função iteradora g_1).
Na vizinhança do ponto fixo $z = 1/\alpha$, o gráfico de g_7 é muito mais "achatado"do que o gráfico de g_1. Isso explica a razão pela qual devemos esperar que as iteradas produzidas usando a função iteradora g_7 se aproximem muito mais rapidamente de z do que no caso de efectuarmos iterações da função g_1.

(e) Fazendo $x_0 = 1/10$ e usando cálculos com precisão de pelo menos 500 dígitos decimais, a quarta e quinta iteradas do método gerado por g_7 são coincidentes, produzindo o número mostrado na Figura 2.25. Podemos por conseguinte garantir que todos os dígitos do número figurado são significativos. Os cálculos foram efectuados no sistema *Mathematica* [38]. ♦

2.7 Leituras aconselhadas

K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & sons, New York, 1978, Ch. 2.

M. M. Graça, *Maps for global separation of roots*, Electronic Transactions on Numerical Analysis (ETNA), Vol. 45, 241-256, 2016. etna.mcs.kent.edu/vol.45.2016/pp241-256.dir/pp241-256.pdf

W. M. Kahan, *Personal calculator has key to solve any equation* $f(x) = 0$, Hewlett-Packard Journal, Vol. 30, 12, Dec. 1979, 20-26.
http://www.hpl.hp.com/hpjournal/pdfs/IssuePDFs/1979-12.pdf.

A. Knoebel, R. Laubenbacher, J. Lodder, D. Pengelley *Mathematical Masterpieces, Further Chronicles by the Explorers*, Springer, 2007, Ch. 2.

Z. Rached, *Arbitrary Order Iterations*, European Int. J. Science and Technology, Vol 2, 5, 191-195, 2013.

P. Sebah, X. Gourdon, *Newton's method and high order iterations*, disponível em http://old.sztaki.hu/~bozoki/oktatas/nemlinearis/SebahGourdon-Newton.pdf

J. Stillwell, *Elements of Algebra, Geometry, Numbers, Equations*, Springer, New York, 2001.

J. Verbeke, R. Cools, *The Newton-Raphson method*, Int. J. Math. Educ. Sc. Tech. 26:2 (2006), 177-193.

Capítulo 3

Métodos numéricos para sistemas de equações

Neste capítulo trataremos de métodos computacionais para a resolução de sistemas de equações (lineares e não lineares). Para a análise do erro destes métodos, necessitaremos frequentemente de recorrer a normas vectoriais e matriciais, pelo que começaremos por fazer uma breve introdução sobre este tema.

3.0.1 Normas matriciais

Seja E um espaço linear. A grandeza de um elemento de E é traduzida numericamente através da norma desse elemento. Tipicamente, nesta disciplina, teremos $E = \mathbb{R}^n$ (vectores de n componentes reais) ou $E = \mathbb{R}^{n\times n}$ (matrizes reais de n linhas e n colunas). Comecemos por relembrar a definição de norma de um elemento de E.

Definição 3.1. Uma aplicação ϕ de E em $\mathbb{R}_0^+$ diz-se uma norma se satisfizer as seguintes condições:

1. $\phi(x) \geq 0, \quad \forall x \in E$, sendo $\phi(x) = 0$ se e só se $x = 0$.
2. $\phi(\lambda\, x) = |\lambda\,|\phi(x), \quad \forall x \in E,\, \lambda \in \mathbb{R}$.
3. $\phi(x + y) \leq \phi(x) + \phi(y), \quad \forall x, y \in E$.

Começamos por rever alguns exemplos de normas em $\mathbb{R}^n$. Como habitualmente, representaremos qualquer elemento de $\mathbb{R}^n$ por $x = (x_1, x_2, \ldots, x_n)$, onde $x_i \in \mathbb{R}$.

Norma do máximo:

$$\phi(x) = \|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \ .$$

Norma 1:

$$\phi(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i| \ .$$

Norma euclidiana:

$$\phi(x) = \|x\|_2 = \sqrt{\sum_i^n |x_i|^2} = \left(\sum_{i=1}^n x_i^2\right)^{1/2} .$$

Norma p:

$$\phi(x) = \|x\|_p = \left(\sum_i^n |x_i|^p\right)^{1/p} , \qquad p \geq 1 .$$

Note-se que a norma 1 e a norma euclidiana são casos particulares das normas p, respectivamente para $p = 1$ e $p = 2$.

Pode provar-se que todos os exemplos anteriores definem normas, isto é, satisfazem as três condições da Definição 3.1. A norma $||x||_\infty$ obtém-se como limite da norma $||x||_p$, quando $p \to \infty$.

Passamos agora a considerar o caso de $E = \mathbb{R}^{n\times n}$. Os elementos de E são matrizes reais, de n linhas e n colunas, isto é, matrizes do tipo $n \times n$. Por exemplo, a matriz

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} .$$

Quando nos referirmos a uma matriz $A \in \mathbb{R}^{n\times n}$, designaremos as entradas de A por a_{ij}.

Represente-se por $\|\cdot\|_v$ uma dada norma em $\mathbb{R}^n$. A partir dessa norma vectorial podemos definir uma norma $\|.\|_M$ em E, da seguinte forma:

Definição 3.2. Seja $A \in \mathbb{R}^{n\times n}$ e $x \in \mathbb{R}^n$.

$$\|A\|_M = \max_{x\in\mathbb{R}^n,\, x\neq 0} \frac{\|A\,x\|_v}{\|x\|_v} . \tag{3.1}$$

Dizemos que a norma $\|\cdot\|_M$ é a norma matricial *induzida pela norma vectorial* $\|.\|_v$.

A Definição 3.2 permite-nos associar uma norma matricial a cada uma das normas vectoriais anteriormente introduzidas.

Propriedades da norma induzida

A norma matricial $||\cdot||_M$ goza de algumas propriedades essenciais, que passamos a referir.

(i) A norma $\|\cdot\|_M$ é *compatível* com a norma $\|\cdot\|_v$, isto é,

$$\|A\,x\|_v \leq \|A\|_M \,\|x\|_v, \quad \forall x \in \mathbb{R}^n, \quad \forall A \in \mathbb{R}^{n\times n}. \tag{3.2}$$

Esta propriedade é uma consequência imediata da fórmula (3.1), e é por vezes referida como propriedade *submultiplicativa* das normas induzidas.

(ii) A norma $\|\cdot\|_M$ é *regular*, isto é,

$$\|A\,B\|_M \leq \|A\|_M \,\|B\|_M, \quad \forall A, B \in \mathbb{R}^{n\times n}. \tag{3.3}$$

Esta propriedade decorre da propriedade submultiplicativa anterior.

(iii) A matriz identidade $I \in \mathbb{R}^{(n\times n)}$ possui norma induzida de valor unitário,

$$||I||_M = 1 \ .$$

Esta propriedade resulta imediatamente da definição dada para norma induzida.

Note que uma generalização possível da norma vectorial euclidiana a matrizes é

$$||A||_F = (\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}^2)^{1/2}, \tag{3.4}$$

A norma (3.4) é conhecida como norma de Frobenius[1] ou de Schur.[2]
Note-se que para a norma $||\,.\,||_F$, se tem $||I||_F = \sqrt{n}$. Conclui-se, portanto, que a norma $||\cdot||_F$, não é uma norma matricial induzida por uma norma vectorial, visto que a norma da matriz identidade é $||I||_F \neq 1$.

Normas usuais

Mostra-se que as normas matriciais dadas a seguir são induzidas pelas normas vectoriais p mais correntes, ou seja, fazendo $p = 1$, $p = 2$ e $p = \infty$ (ver, por exemplo, [24], p. 34).

1. A norma matricial induzida pela *norma do máximo*, isto é, para $p = \infty$, chama-se *norma por linha*,

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \ . \tag{3.5}$$

[1] Ferdinand Georg Frobenius, 1849 -1917, matemático alemão.
[2] Issai Schur, 1875 - 1941, matemático nascido na Bielorrússia, professor na Alemanha.

2. A norma matricial induzida pela norma vectorial 1 chama-se *norma por coluna*. É definida pela fórmula

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \ . \tag{3.6}$$

3. Prova-se que a norma matricial induzida pela norma (vectorial) euclidiana ($p = 2$) é

$$\|A\|_2 = \sqrt{\rho(A^T A)}, \tag{3.7}$$

onde A^T designa a matriz transposta de A e o símbolo $\rho(M)$ representa o *raio espectral* da matriz M, que se define como o máximo dos módulos dos valores próprios de M, ou seja,

Definição 3.3. Sejam $\lambda_1, \lambda_2, \ldots, \lambda_n$ os valores próprios da matriz $A \in \mathbb{R}^{n\times n}$. Define-se raio espectral de A por

$$\rho(A) = \max_{i=1,\ldots,n} |\lambda_i| \ . \tag{3.8}$$

Note-se que, se A for uma matriz simétrica, isto é, se $A^T = A$, são válidas as igualdades

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \rho(A) \ . \tag{3.9}$$

Isto é, para matrizes simétricas A, a norma $||A||_2$ coincide com o seu raio espectral. Retenha-se a este propósito a ideia de que o raio espectral de uma matriz está intimamente ligado ao seu "comprimento" ou grandeza. Como se verá mais adiante, matrizes cujo raio espectral seja inferior à unidade revestem-se de interesse muito particular.

Exemplo 3.1. *Pretende-se determinar as normas usuais da matriz*

$$A = \begin{bmatrix} 2 & 1 & -3 \\ 1 & 3 & 4 \\ 2 & -1 & 3 \end{bmatrix}.$$

As normas matriciais induzidas anteriormente referidas, dão-nos

$$\|A\|_\infty = \max(6, 8, 6) = 8,$$

e

$$\|A\|_1 = \max(5, 5, 10) = 10 \ .$$

Para se calcular $\|A\|_2$ é necessário começar por determinar a matriz (simétrica) $B = A^T A$ a seguir,

$$B = A^T A = \begin{bmatrix} 9 & 3 & 4 \\ 3 & 11 & 6 \\ 4 & 6 & 34 \end{bmatrix} .$$

Os valores próprios de B são, aproximadamente, $\lambda_1 = 6.8$, $\lambda_2 = 10.9$ e $\lambda_3 = 36.3$. Logo, $\rho(A^T A) = 36.3$ e

$$||A||_2 \simeq \sqrt{36.3} \simeq 6.02 .$$

Interessa comparar o raio espectral da matriz A com a respectiva norma $||A||_2$. Os valores próprios de A são o número real $\lambda_1 = 3.69$, e os dois números complexos conjugados $\lambda_{2,3} = 2.15 \pm i\,3.07$, donde $|\lambda_2| = |\lambda_3| \simeq 3.75$. Por conseguinte, $\rho(A) \simeq 3.75$, e

$$\rho(A) \leq ||A||_2 .$$

♦

Passamos a designar a norma matricial induzida pela norma vectorial $||\cdot||_p$ por $||A||_p$. No anterior Exemplo 3.1, qualquer das normas de A é maior que o raio espectral da matriz. Tal não acontece por acaso, conforme é mostrado a seguir.

Teorema 3.1. Seja $A \in \mathbb{R}^{n\times n}$. Qualquer que seja a norma matricial $||\cdot||_M$, induzida pela norma vectorial $||\cdot||_V$ em $\mathbb{R}^n$, é válida a desigualdade

$$\rho(A) \leq \|A\|_M, \quad \forall A \in \mathbb{R}^{n\times n} \tag{3.10}$$

Demonstração. Seja $x \neq 0$ um vector próprio de A associado ao valor próprio λ, tal que $|\lambda| = \rho(A)$. Logo,

$$\|A\,x\|_V = \|\lambda\,x\|_V = |\lambda|\,\|x\|_V . \tag{3.11}$$

Assim,

$$\|A\|_M = \max_{x\in\mathbb{R}^n, x\neq 0} \frac{\|A\,x\|_V}{\|x\|_V} \geq |\lambda| = \rho(A), \tag{3.12}$$

donde resulta a afirmação do teorema. □

Uma vez que geralmente é mais fácil calcular a norma de uma matriz do que o seu raio espectral, a relação (3.10) será frequentemente invocada.

3.1 Condicionamento de sistemas lineares

Como vimos no Capítulo 1, um dos aspectos importantes a ter em consideração quando se analisam métodos numéricos para aproximar a solução de um determinado problema é a sensibilidade desses métodos em relação a pequenos erros nos dados. Se for dado um certo sistema linear,

$$A\,x = b,$$

tal que $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^{n\times 1}$, os dados são o segundo membro b do sistema e a matriz A dos coeficientes, que supomos ser não singular (ou seja, invertível).

Vamos analisar até que ponto um pequeno erro, em termos relativos, do vector b ou da matriz A, pode afectar a solução do sistema.

Representemos por $\bar{A}$ uma matriz *perturbada*, ou seja, uma matriz distinta de A mas "próxima",

$$\bar{A} \approx A.$$

Analogamente, representemos por $\bar{b}$ um vector que resulta de uma perturbação do segundo membro do sistema,

$$\bar{b} \approx b.$$

Se substituirmos A por $\bar{A}$ e b por $\bar{b}$ no sistema inicial, obteremos um novo sistema, cuja solução representaremos por $\bar{x}$.

Vamos designar por *erro relativo de um vector* $\bar{x}$ (numa certa norma vectorial V) o quociente

$$\|\delta_{\bar{x}}\|_V = \frac{\|\bar{x} - x\|_V}{\|x\|_V} \ . \tag{3.13}$$

Analogamente, designaremos por *erro relativo de uma matriz* $\bar{A}$ (na norma matricial induzida), o quociente

$$\|\delta_{\bar{A}}\|_M = \frac{\|\bar{A} - A\|_M}{\|A\|_M} \ . \tag{3.14}$$

Escolhida uma certa norma vectorial e a respectiva norma matricial induzida, é nosso objectivo estimar o erro relativo $\|\delta_{\bar{x}}\|_V$, em função dos erros relativos $\|\delta_{\bar{b}}\|_V$ e $\|\delta_{\bar{A}}\|_M$.

Generalizando noção análoga para funções (ver parágrafo 1.2.2, pág. 20), comecemos por definir o que se entende por condicionamento de um sistema linear.

Definição 3.4. Um sistema linear não singular diz-se *bem condicionado* se e só se, a pequenos erros relativos do segundo membro e/ou da matriz dos coeficientes correspondem pequenos erros relativos na solução.

3.1.1 Perturbações do segundo membro

Para analisarmos o problema do condicionamento, comecemos por considerar o caso mais simples em que a matriz $A = \bar{A}$, ou seja, $\|\delta_{\bar{A}}\|_p = 0$. Nesse caso, temos

$$A\,\bar{x} = \bar{b} \ . \tag{3.15}$$

Usando (3.15) obtém-se

$$\bar{x} - x \ = \ A^{-1}\,(\bar{b} - b) \ .$$

Por conseguinte, atendendo a (3.2), qualquer que seja a norma vectorial escolhida, é válida a seguinte estimativa para o erro absoluto de $\bar{x}$:

$$\|\bar{x} - x\|_V \leq \|A^{-1}\|_M \, \|\bar{b} - b\|_V \ . \tag{3.16}$$

Usando de novo (3.2), da igualdade $A\,x = b$, obtém-se

$$\|b\|_V \leq \|A\|_M \, \|x\|_V .$$

Portanto,

$$\frac{1}{\|x\|_V} \leq \frac{\|A\|_M}{\|b\|_V}, \quad \text{para} \quad x, b \neq 0 \ . \tag{3.17}$$

Uma vez subentendido qual a norma vectorial e correspondente norma matricial em uso, podemos ignorar os símbolos $\cdot_V$ e $\cdot_M$ em (3.17).

Multiplicando cada um dos membros de (3.16) pelo membro correspondente de (3.17), resulta

$$\frac{||x - \bar{x}||}{||x||} \leq ||A||\,||A^{-1}||\,\frac{||b - \bar{b}||}{||b||} \ . \tag{3.18}$$

Obtivemos assim a estimativa que procurávamos para o erro relativo na solução, em função do erro relativo do segundo membro.

A presença do factor $||A||\,||A^{-1}||$ na desigualdade (3.18) sugere-nos a definição a seguir.

Definição 3.5. Seja A uma matriz invertível. Chama-se *número de condição de* A (na norma M) ao valor

$$cond_M(A) = \|A\|_M \, \|A^{-1}\|_M . \tag{3.19}$$

De agora em diante vamos supor que as normas em jogo são as normas p usuais.

Uma relação entre o erro relativo da solução de um sistema linear e o erro relativo do seu segundo membro é dada pela desigualdade (3.18). Assim, *se o número de condição de A for elevado*, pode resultar que pequenos erros relativos do segundo membro provoquem erros muito maiores na solução — uma situação que, atendendo à Definição 3.4, significará que o sistema possui a propriedade indesejável de ser mal condicionado.

Note-se que o número de condição de uma matriz é sempre maior ou igual a 1, desde que consideremos normas matriciais induzidas.
Com efeito, como

$$I = A\,A^{-1},$$

resulta

$$1 = ||I|| \leq ||A||\,||A^{-1}|| = cond(A) \ .$$

Por conseguinte, um sistema bem condicionado é aquele que possua um número de condição não muito maior que 1.

Número de condição espectral

Uma definição alternativa do número de condição utiliza o raio espectral,

$$cond_*(A) = \rho(A) \times \rho(A^{-1}) \; . \tag{3.20}$$

De acordo com o Teorema 3.1, pág. 97, podemos escrever

$$cond_*(A) \leq cond_p(A), \tag{3.21}$$

qualquer que seja a norma matricial p considerada ($p \geq 1$). Daqui resulta que se o número de condição $cond_*(A)$ for elevado, todos os números de condição da matriz são elevados, pelo que o sistema é mal condicionado. No entanto, pode acontecer que o sistema seja mal condicionado mesmo que o número de condição $cond_*(A)$ seja pequeno.

Atendendo a que os valores próprios[3] de A^{-1} são os inversos dos valores próprios de A, o número de condição $cond_*(A)$ pode escrever-se sob a forma

$$cond_*(A) = \frac{\max_{\lambda_i \in \sigma(A)} |\lambda_i|}{\min_{\lambda_i \in \sigma(A)} |\lambda_i|} \; . \tag{3.22}$$

No caso de a matriz A ser *simétrica*, como foi observado antes, a sua norma euclidiana coincide com o raio espectral, pelo que podemos escrever,

$$cond_2(A) = cond_*(A) \; . \tag{3.23}$$

3.1.2 Perturbação da matriz e do segundo membro

Vejamos agora o caso geral em que o sistema linear pode estar afectado de erros, não só no segundo membro b, mas também na própria matriz A.

Teorema 3.2. Consideremos o sistema linear $A\,x = b$, onde A é uma matriz invertível. Sejam $\delta_{\bar{A}}$ e $\delta_{\bar{b}}$ definidos respectivamente pelas igualdades (3.14) e (3.13), e suponhamos que

$$\|A - \bar{A}\| \leq \frac{1}{\|A^{-1}\|} \; . \tag{3.24}$$

É satisfeita a desigualdade

$$\|\delta_{\bar{x}}\| \leq \frac{cond(A)}{1 - cond(A)\,\|\delta_{\bar{A}}\|} \left(\|\delta_{\bar{A}}\| + \|\delta_{\bar{b}}\|\right). \tag{3.25}$$

[3] O conjunto dos valores próprios de uma matriz A, ou seja o espectro de A, será denotado por $\sigma(A)$ ou $Sp(A)$.

Demonstração. Ver, por exemplo, [1]. □

Observação. Note-se que a desigualdade (3.18), pág. 99, é um caso particular de (3.25), que se obtém fazendo $\|\delta_{\bar{A}}\|_p = 0$.

A desigualdade (3.25) confirma a conclusão de que os sistemas lineares com números de condição elevados são mal condicionados. O exemplo que se segue mostra como os problemas de mau condicionamento podem surgir mesmo em sistemas de pequenas dimensões, e com matrizes aparentemente "bem comportadas".

Exemplo 3.2. *Consideremos o sistema linear $A\,x = b$, onde*

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}, \quad b = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}. \tag{3.26}$$

Mostre-se que o sistema é mal condicionado, efectuando pequenas perturbações quer do segundo membro b, quer da própria matriz A.

Verifica-se imediatamente que a solução deste sistema é $x = (1, 1, 1, 1)^T$[4]. A matriz A é simétrica e não singular[5]. A sua norma (por linhas ou por colunas) é

$$\|A\|_\infty = \|A\|_1 = max(32, 23, 33, 31) = 33\ .$$

Se substituirmos o vector b pelo vector $\bar{b}$, seja

$$\bar{b} = (32.1, 22.9, 33.1, 30.9)^T,$$

a solução do sistema passa a ser

$$\bar{x} = (9.2, -12.6, 4.5, -1.1)^T,$$

a qual é muito diferente da solução do sistema inicial. Por outras palavras, uma perturbação relativa do segundo membro,

$$\|\delta_{\bar{b}}\|_\infty = \frac{0.1}{33} \approx 0,3\%,$$

leva-nos a uma nova solução, cuja norma $||\bar{x}||_\infty$ é cerca de 13 vezes maior que a da solução original.

[4]Note que cada entrada de b tem o valor da soma das entradas da linha correspondente da matriz A.

[5]Pode verificar que $det(A) = 1 \neq 0$.

Observemos ainda o que acontece se a matriz A sofrer uma ligeira perturbação das suas entradas, sendo substituída por

$$\bar{A} = \begin{bmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 5 & 9 & 9.98 \end{bmatrix}, \tag{3.27}$$

mantendo-se o segundo membro inalterado. Neste caso, a solução do sistema passa a ser

$$\bar{x} = (-81, 137, -34, 22)^T.$$

Verifica-se que a diferença em relação à solução inicial é ainda mais acentuada. Entretanto, a norma da perturbação é relativamente pequena, pois

$$\|A - \bar{A}\|_\infty = max(0.3, 0.12, 0.13, 0.03) = 0.3,$$

donde

$$\|\delta_{\bar{A}}\|_\infty = \frac{0.3}{33} \approx 0,9\,\%.$$

Vejamos como interpretar estes factos com base na teoria que expusemos previamente. Para o efeito, precisamos de conhecer a inversa de A,

$$A^{-1} = \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}. \tag{3.28}$$

Podemos imediatamente constatar que

$$\|A^{-1}\|_\infty = max(82, 136, 35, 21) = 136\ .$$

Assim, o número de condição de A, na norma ∞ (que coincide com o número de condição na norma 1, pois a matriz A^{-1} é simétrica), tem o valor

$$cond_\infty(A) = cond_1(A) = 33 \times 136 = 4488\ .$$

Conhecendo o valor do número de condição, já não nos surpreende o facto de as pequenas perturbações que introduzimos no segundo membro e na matriz terem alterado completamente a solução. Com efeito, a estimativa (3.18), pág. 99, aplicada a este caso, diz-nos que

$$||\delta_{\bar{x}}|| \leq 4488 \times 0.3\ \% = 1346\ \%,$$

o que explica inteiramente os maus resultados obtidos, no que diz respeito à perturbação do segundo membro do sistema.

Note-se que para o caso em que se perturbou a a matriz A não se pode aplicar a estimativa (3.25), pág. 100, uma vez que, para a perturbação considerada, não é satisfeita a condição

$$\|A - \bar{A}\|_\infty \leq \frac{1}{\|A^{-1}\|_\infty}.$$

No entanto, dado o elevado valor do número de condição obtido, é expectável que a solução do sistema sofra grandes alterações quando se perturbam os dados.

⧫

Deixamos ao leitor a resolução das questões que constam dos dois exercícios a seguir.

Exercício 3.1. *Seja A uma matriz quadrada, de dimensão $n \times n$, com a forma*

$$A = \begin{bmatrix} 1 & -1 & \dots & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -1 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}.$$

1. *Calcule A^{-1} e determine os números de condição $cond_1(A)$ e $cond_\infty(A)$.*
2. *Sejam b_1 e b_2 dois vectores de $\mathbb{R}^n$ tais que*

$$\|\delta_b\|_\infty = \frac{\|b_1 - b_2\|_\infty}{\|b_1\|_\infty} \leq 10^{-5},$$

e x_1 e x_2 as soluções dos sistemas $A\,x = b_1$ e $A\,x = b_2$. Determine um majorante de

$$\|\delta_x\|_\infty = \frac{\|x_1 - x_2\|_\infty}{\|x_1\|_\infty},$$

para $n = 20$. Comente quanto ao condicionamento de um sistema arbitrário $A\,x = b$.

⧫

Exercício 3.2. *Seja $a \neq 3 \in \mathbb{R}$ e*

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ a & 0 & 3 \end{bmatrix},$$

Suponhamos que, ao resolver o sistema $A\,x = b$, com um certo valor de a, se obteve a solução $\bar{x} = (1, 1, 1)$.

Admitindo que o valor de a está afectado de um certo erro, de valor absoluto não superior a um uma certa tolerância ϵ, determine um majorante de $\|\Delta\,\bar{x}\|_\infty$, onde $\Delta\,\bar{x}$ é a diferença entre a solução obtida e a que se obteria se fosse conhecido o valor exacto de a.

⧫

3.2 Métodos directos para sistemas lineares *

Para resolver um sistema linear, podemos considerar as duas alternativas a seguir.[6]

1. Reduzir o sistema a uma forma mais simples, de modo a obter a solução exacta através de substituições adequadas. Nesse caso, dizemos que estamos a aplicar um *método directo*.

2. Determinar a "solução"por um método de aproximações sucessivas, utilizando um *método iterativo*.

Começaremos por discutir alguns métodos directos. Quando se utiliza métodos deste tipo e cálculos exactos, sabe-se que o seu erro é nulo, visto que o método (teoricamente) conduz à solução exacta do sistema. Porém, tal não significa que a solução obtida através de uma máquina seja exacta, uma vez que ao efectuarmos cálculos numéricos são inevitáveis os erros de arredondamento.

3.2.1 Método de eliminação de Gauss

Um dos métodos mais simples para a resolução de sistemas lineares é o *método da eliminação de Gauss*[7].

A ideia básica deste método consiste em reduzir o sistema dado, $A\,x = b$ (com A quadrada), a um sistema equivalente, $U\,x = b'$, onde U é uma matriz triangular superior. Este último sistema pode ser resolvido por substituição ascendente ou regressiva.

Assim, podemos dizer que a resolução de um sistema pelo método de Gauss se divide em três etapas:

1. Redução da matriz A à forma triangular superior.

2. Transformação do segundo membro do sistema.

3. Resolução do sistema a partir da matriz triangular superior obtida em 1.

Vejamos com mais pormenor em que consiste cada uma destas etapas e avaliemos, em termo de número de operações aritméticas, o volume dos cálculos correspondentes.

[6]Embora os métodos directos para sistemas de equações lineares não constem para avaliação na disciplina de Matemática Computacional, sugere-se ao aluno que assimile os algoritmos versados nesta secção, porquanto eles são fundamentais na bagagem de conhecimentos de um futuro engenheiro.

[7]Johann Carl Friedrich Gauss, 1777 -1855, matemático alemão considerado um dos maiores génios de todos os tempos.

1. Redução da matriz A à forma triangular superior

Suponhamos que a matriz dada é da forma

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Admitindo que $a_{11} \neq 0$, eliminam-se as restantes entradas da primeira coluna de A (começando com a_{21}) somando a cada linha um múltiplo da primeira. Assim, resulta uma nova matriz $A^{(1)}$, da forma

$$A^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)}. \end{bmatrix}.$$

As entradas de $A^{(1)}$ obtêm-se através das relações

$$a_{ij}^{(1)} = a_{ij} - m_{i1}\, a_{1j}, \tag{3.29}$$

onde

$$m_{i1} = \frac{a_{i1}}{a_{11}}\ . \tag{3.30}$$

Ignorando a primeira linha de $A^{(1)}$, repetimos o processo anterior, eliminando as entradas da segunda coluna, abaixo de $a_{22}^{(1)}$.

Repetindo sucessivamente estas transformações, obtém-se em cada passo uma matriz $A^{(k)}$ da forma

$$A^{(k)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nk}^{(k-1)} & \dots & a_{nn}^{(k-1)} \end{bmatrix}.$$

As entradas de $A^{(k)}$ obtêm-se a partir das de $A^{(k-1)}$, através das expressões[8],

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \quad i = k+1 : n, \qquad j = k+1 : n, \tag{3.31}$$

[8]Relembre-se que notação do tipo $i = m : n$, significa $i = m, m+1, \dots, n$.

onde

$$m_{ik} \;=\; \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \tag{3.32}$$

(pressupõe-se que $a_{kk}^{(k-1)} \neq 0$). Ao fim de $n-1$ transformações, obtém-se

$$A^{(n-1)} \;=\; \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & a_{nn}^{(n-1)} \end{bmatrix}. \tag{3.33}$$

No caso de alguma das entradas $a_{kk}^{(k-1)}$ ser igual a zero, torna-se necessário alterar a ordem das linhas. Esse caso será analisado em detalhe mais adiante, durante a resolução do Exemplo 3.3, pág. 108.

Note-se que se a matriz A for não singular, existe sempre uma permutação das suas linhas, de tal forma que A pode ser reduzida à forma (3.33), com todos os elementos da diagonal principal diferentes de zero.

2. Transformação do segundo membro

O segundo membro do sistema $A\,x = b$ é sujeito às mesmas transformações que se efectuaram sobre A, de modo a garantir a equivalência do sistema resultante ao inicial.
Assim, a transformação do vector b também se realiza em $n-1$ passos, sendo a primeira transformada, $b^{(1)}$, obtida segundo a fórmula

$$b_i^{(1)} \;=\; b_i \;-\; m_{i1}\, b_1, \qquad i = 2:n. \tag{3.34}$$

Analogamente, a k-ésima transformada do segundo membro passa a ser,

$$b_i^{(k)} \;=\; b_i \;-\; m_{ik}\, b_k^{(k-1)}, \qquad i = k+1:n. \tag{3.35}$$

Os coeficientes m_{ik} são dados pelas fórmulas (3.30) e (3.32).

3. Resolução do sistema triangular superior

Depois de reduzido o sistema inicial à forma triangular superior, de matriz dada por (3.33), a solução obtém-se facilmente mediante o seguinte processo de subs-

tituições regressivas (ou ascendentes),

$$\begin{aligned} x_n &= \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}} \\ x_{n-1} &= \frac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}\, x_n}{a_{n-1,n-1}^{(n-2)}} \\ &\vdots \\ x_1 &= \frac{b_1 - \sum_{i=2}^{n} a_{1,i}\, x_i}{a_{1,1}} \ . \end{aligned} \tag{3.36}$$

3.2.2 Contagem de operações

Vejamos agora como estimar o número de operações aritméticas necessárias para efectuar cada uma das etapas que acabámos de descrever.

1. Redução da matriz A à forma triangular superior

O número de operações necessárias para a transformação da matriz A está relacionado com o número de vezes que são aplicadas as fórmulas (3.29) e (3.31).

No 1º passo, a fórmula (3.29) é aplicada $(n-1)^2$ vezes. Isto implica que se realizem $(n-1)^2$ multiplicações e outras tantas adições (ou subtracções). Para calcular os quocientes da fórmula (3.30), efectuam-se $n-1$ divisões. Todas estas operações continuam a efectuar-se nos passos seguintes da transformação, mas em menor número, de acordo com o número de entradas que são alteradas em cada passo. Em geral, no k-ésimo passo efectuam-se $(n-k)^2$ multiplicações e outras tantas adições (ou subtracções), assim como $n-k$ divisões.

Assim, o número total de multiplicações $M(n)$ efectuadas na transformação da matriz A, é igual ao número de adições (ou subtracções), $AS(n)$, ou seja,

$$M(n) \ = \ AS(n) \ = \ \sum_{k=1}^{n-1} (n-k)^2 \ = \ \frac{n(n-1)(2n-1)}{6} \ . \tag{3.37}$$

Quanto ao número de divisões, $D(n)$, obtém-se,

$$D(n) \ = \ \sum_{k=1}^{n-1} (n-k) \ = \ \frac{n(n-1)}{2} \ . \tag{3.38}$$

O número total de operações, $TO(n)$, efectuadas na transformação da matriz A é, em termos assimptóticos (ou seja, para valores elevados de n),

$$TO(n) \ = M(n) + AS(n) + D(n) \approx \frac{2\,n^3}{3} + \mathcal{O}(n^2) \ . \tag{3.39}$$

2. Transformação do segundo membro

Quando transformamos o vector b, usamos a fórmula (3.35). No k-ésimo passo do método a fórmula (3.35) é aplicada $n-k$, o que implica $n-k$ multiplicações e outras tantas adições (ou subtracções). Assim, o número total de multiplicações $M(n)$ é igual ao número de adições (ou subtracções), ou seja,

$$M(n) \,=\, AS(n) \,=\, \sum_{k=1}^{n-1}(n-k) \,=\, \frac{n(n-1)}{2}\ . \tag{3.40}$$

Por conseguinte, o número total de operações exigidas na transformação do segundo membro é, em termos assimptóticos,

$$TO(n) \,=\, M(n) \,+\, AS(n) \,\approx\, n^2. \tag{3.41}$$

3. Resolução do sistema triangular

Para resolver o sistema triangular anteriormente obtido, efectuamos as substituições (3.36). Como resulta destas fórmulas, o número total de multiplicações para resolver o sistema é $n(n-1)/2$, igual ao número total de adições (ou subtracções). Quanto ao número de divisões, $D(n)$, é igual a n.

Por conseguinte, o número total de operações efectuadas para resolver o sistema triangular é, em termos assimptóticos,

$$TO(n) \,=\, M(n) \,+\, AS(n) \,+\, D(n) \,\approx\, n^2. \tag{3.42}$$

O maior esforço computacional é efectuado na etapa da triangularização da matriz A, conforme se conclui se compararmos (3.39) com (3.42). Por este motivo, podemos dizer que o número N de operações envolvidas no cálculo da solução do sistema $A\,x = b$ pelo método de eliminação de Gauss é

$$N = \mathcal{O}\left(\frac{2}{3}n^3\right).$$

Exemplo 3.3. *Consideremos o sistema linear $A\,x = b$, onde*

$$A = \begin{bmatrix} 2 & 1 & 3 \\ -2 & -1 & 1 \\ 2 & 4 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix}. \tag{3.43}$$

Pretende-se resolver este sistema pelo método da eliminação de Gauss.

Comecemos por reduzir A à forma triangular superior. O primeiro passo consiste em transformar a matriz A na matriz $A^{(1)}$. Usando as fórmulas (3.29) e (3.30), obtém-se:

$$\begin{aligned} a_{22}^{(1)} \,&=\, a_{22} \,-\, m_{21}\,a_{12} = 0 \\ a_{23}^{(1)} \,&=\, a_{23} \,-\, m_{21}\,a_{13} = 4, \end{aligned}$$

onde

$$m_{21} = \frac{a_{21}}{a_{11}} = -1 \ .$$

Verifica-se que o novo elemento da diagonal principal, $a_{22}^{(1)}$, é nulo. Como sabemos, neste caso não é possível aplicar o método da eliminação de Gauss sem proceder a uma troca de linhas – mais precisamente, troquemos a segunda linha com a terceira. Obtém-se assim o novo sistema $A'\,x = b'$, onde

$$A' = \begin{bmatrix} 2 & 1 & 3 \\ 2 & 4 & 2 \\ -2 & -1 & 1 \end{bmatrix}, \qquad b' = \begin{bmatrix} 5 \\ 4 \\ -1 \end{bmatrix}.$$

Aplicando o método da eliminação de Gauss a este sistema, usemos de novo as fórmulas (3.29) e (3.30):

$$\begin{array}{rcl} a_{22}^{(1)'} & = & a'_{22} - m'_{21}\, a_{12} = 4 - 1 = 3 \\ a_{23}^{(1)'} & = & a'_{23} - m'_{21}\, a_{13} = 2 - 3 = -1 \\ a_{32}^{(1)'} & = & a'_{32} - m'_{31}\, a_{12} = -1 + 1 = 0 \\ a_{33}^{(1)'} & = & a'_{33} - m'_{31}\, a_{13} = 1 + 3 = 4, \end{array}$$

onde

$$m'_{21} = \frac{a'_{21}}{a_{11}} = 1,$$

$$m'_{31} = \frac{a'_{31}}{a_{11}} = -1 \ .$$

Resulta assim a matriz triangular superior

$$A' = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 4 \end{bmatrix}.$$

A segunda etapa da aplicação do método da eliminação de Gauss consiste em transformar o segundo membro do sistema, isto é, o vector b'. Para isso, utilizamos a fórmula (3.34), que neste caso nos dá

$$\begin{array}{rcl} b_2^{(1)'} & = & b'_2 - m'_{21}\, b'_1 = 4 - 5 = -1 \\ b_3^{(1)'} & = & b'_3 - m'_{31}\, b'_1 = -1 + 5 = 4 \ . \end{array}$$

Obtemos assim o vector transformado $b^{(1)'} = (5, -1, 4)^T$.

Por último, resta-nos resolver o sistema triangular superior $A^{(1)'}\,x = b^{(1)'}$. Para isso, usamos substituições ascendentes, isto é, começamos por determinar x_3 a partir da última equação, para depois determinar x_2 da segunda e x_1 da primeira. Usando as fórmulas (3.36), obtém-se

$$
\begin{aligned}
x_3 &= \frac{b_3^{(1)'}}{a_{33}^{(1)}} = 1 \\
x_2 &= \frac{b_2^{(2)} - a_{23}^{(1)}\, x_3}{a_{22}^{(1)}} = \frac{-1+1}{2} = 0 \\
x_1 &= \frac{b_1 - a_{13}\, x_3 - -a_{12}\, x_2}{a_{1,1}} = \frac{5-3}{2} = 1 \ .
\end{aligned}
$$

Pelo que a solução do sistema é $x = (1, 0, 1)^T$. ♦

3.2.3 Influência dos erros de arredondamento

Ao relembrarmos o método de eliminação de Gauss no parágrafo anterior, não entrámos em consideração com os erros cometidos durante os cálculos. Na Secção 3.1, pág. 97, já vimos que pequenos erros nos dados iniciais do sistema podem afectar muito a solução, caso a matriz seja mal condicionada. Com efeito, além dos erros dos dados iniciais, há que ter em conta também o erro computacional, resultante dos arredondamentos efectuados durante os cálculos.

Um dos inconvenientes do método de Gauss, assim como de outros métodos directos de que falaremos adiante, consiste em que esses erros têm frequentemente tendência para se propagar durante os cálculos, de tal modo que podem adquirir um peso muito grande na solução, mesmo que o sistema seja bem condicionado. No entanto, o efeito destes erros pode ser bastante atenuado se durante os cálculos forem usadas precauções adequadas, como a chamada *estratégia de pivot* de que nos ocuparemos a seguir.

Ao discutirmos a transformação da matriz A, vimos que é necessário que todos os elementos da diagonal principal da matriz triangular superior U sejam diferentes de 0. Estes elementos foram representados por $a_{kk}^{(k-1)}$ e são designados geralmente como *pivots*, dada a sua importância para a aplicação do método de Gauss[9].

Vimos também que, no caso de um dos pivots ser nulo, se podia mesmo assim aplicar o método desde que se efectuasse uma troca de linhas na matriz.
Se o pivot não for nulo, mas próximo de 0, o método continua a ser teoricamente aplicável, mesmo sem trocas de linhas. Só que, ao ficarmos com um denominador muito pequeno no segundo membro de (3.32), pág. 106, cria-se uma situação em que os erros de arredondamento podem propagar-se de uma forma desastrosa. A estratégia de pivot tem por objectivo evitar que isto aconteça. Para esse efeito, em cada passo da transformação da matriz, verifica-se a grandeza do pivot e,

[9]Em língua francesa *pivot* tem o significado de *base, apoio.*

caso se considere conveniente, efectua-se uma troca de linhas que nos permita substituir o pivot inicial por outro de maior grandeza.

A referida estratégia de pivot possui diversas variantes, sendo aqui apenas abordadas a *pesquisa parcial* e a *pesquisa total* de pivot.

Pesquisa parcial de pivot

Em cada passo da transformação da matriz A é inspeccionada a coluna k da matriz $A^{(k-1)}$ (ver expressão (3.31), pág. 105), mais precisamente, as entradas (ou componentes) dessa coluna que se situam abaixo da diagonal principal. Seja

$$c_k = \max_{k \leq i \leq n} \left| a_{ik}^{(k-1)} \right| . \tag{3.44}$$

Se o máximo no segundo membro de (3.44) for atingido para $i = k$, isso significa que o actual pivot é, em módulo, a maior entrada daquela coluna. Nesse caso, continuam-se os cálculos normalmente. Se o máximo for atingido para um certo $i \neq k$, então troca-se a linha k com a linha i e só depois se prosseguem os cálculos. Evidentemente, ao fazer esssa troca, também se efectua uma permutação correspondente nas entradas do vector b.

Pesquisa total de pivot

De acordo com esta estratégia, é inspeccionada não só a coluna k da matriz $A^{(k-1)}$, mas também todas as colunas subsequentes. Seja

$$c_k = \max_{k \leq i,j \leq n} \left| a_{ij}^{(k-1)} \right| . \tag{3.45}$$

Sejam i' e j', respectivamente, os valores dos índices i e j para os quais é atingido o máximo no segundo membro de (3.45). Se i' não coincidir com k, a linha i' troca de lugar com a linha k. Se, além disso, j' não coincidir com k, então a coluna j' também vai trocar de lugar com a coluna k (o que corresponde a uma troca de posição das incógnitas $x_{j'}$ e x_k).

Comparando as duas variantes de pesquisa de pivot, conclui-se que a pesquisa total é bastante mais dispendiosa do que a parcial, uma vez que exige um número de comparações muito maior.

A prática do cálculo numérico tem demonstrado que, na grande maioria dos casos, a pesquisa parcial conduz a resultados praticamente tão bons como os da total. Isto explica por que razão a pesquisa parcial seja mais frequentemente escolhida quando se elaboram algoritmos baseados no método de Gauss.

O exemplo que se segue mostra até que ponto os erros de arredondamento podem influir na solução de um sistema linear, quando é aplicado o método da eliminação de Gauss. Vamos observar como a pesquisa parcial de pivot pode contribuir para melhorar esta situação.

Exemplo 3.4. *Pretende-se aplicar o método de eliminação de Gauss para calcular a solução do sistema linear* $A\,x = b$*, onde*

$$A = \begin{bmatrix} 10^{-6} & 0 & 1 \\ 1 & 10^{-6} & 2 \\ 1 & 2 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}. \tag{3.46}$$

Os cálculos serão efectuados com arredondamento simétrico para 6 dígitos na mantissa. Comparamos a solução, sem e com pesquisa parcial de pivot[10]*.*

Ao resolver o sistema utilizando o método da eliminação de Gauss, chegamos ao sistema equivalente $U\,x = b'$, onde

$$U = \begin{bmatrix} 10^{-6} & 0 & 1 \\ 0 & 10^{-6} & 2-10^{6} \\ 0 & 0 & 2\times 10^{12} - 5\times 10^{6} - 1 \end{bmatrix}, \quad b' = \begin{bmatrix} 1 \\ 3-10^{6} \\ 2\times 10^{12} - 7\times 10^{6} + 2 \end{bmatrix}. \tag{3.47}$$

Suponhamos que os cálculos são efectuados num computador em que os números são representados, no sistema decimal, com *seis dígitos* na mantissa. Em vez de U e b', tem-se[11]

$$\begin{aligned} \tilde{U} &= \begin{bmatrix} 1.00000\times 10^{-6} & 0 & 1.00000 \\ 0 & 1.00000\times 10^{-6} & -0.999998\times 10^{6} \\ 0 & 0 & 1.99999\times 10^{12} \end{bmatrix} \\ \tilde{b} &= \begin{bmatrix} 1 \\ -0.999997\times 10^{6} \\ 1.99999\times 10^{12} \end{bmatrix}. \end{aligned} \tag{3.48}$$

Assim, ao resolvermos o sistema (3.48) por substituições regressivas, obtemos

$$\begin{aligned} \tilde{x_3} &= \frac{1.99999\times 10^{12}}{1.99999\times 10^{12}} = 1.00000 \\ \tilde{x_2} &= \frac{-0.999997\times 10^{6} + 0.999998\times 10^{6}\,\tilde{x_3}}{1.00000\times 10^{-6}} = 1.00000\times 10^{6} \\ \tilde{x_1} &= \frac{1.00000 - 1.00000\,\tilde{x_3}}{1.00000\times 10^{-6}} = 0\ . \end{aligned}$$

[10] Pode verificar que $det(A) \simeq 2$, pelo que o sistema é não singular. Note que sistemas *quase singulares*, isto é, cuja matriz possui determinante próximo de 0, são de evitar porquanto o seu número de condição é geralmente muito grande. No caso de sistemas quase singulares mesmo a pesquisa de pivot não permite em geral contrariar a instabilidade numérica associada a sistemas dessa natureza.

[11] As entradas de $\tilde{U}$ e $\tilde{b}$ poderiam ser escritas usando a notação de ponto flutuante introduzida no Capítulo 1, mas aqui preferimos apresentar os resultados na forma utilizada habitualmente nas máquinas de calcular vulgares.

Substituindo os valores calculados no sistema dado, verifica-se que eles estão longe de o satisfazer, o que indica que este resultado apresenta um erro relativo muito grande. Este erro, no entanto, não tem a ver com o condicionamento do sistema visto que o número de condição da matriz A tem o valor

$$cond_\infty(A) \;=\; \|A\|_\infty \|A^{-1}\|_\infty \approx 3 \times 4 = 12,$$

pelo que o sistema não se pode considerar mal condicionado. Há portanto razões para se suspeitar que o mau resultado obtido resulta da instabilidade numérica do método, a qual, como vimos, pode ser contrariada através da pesquisa de pivot.

Vejamos que resultado obtemos aplicando *pesquisa parcial de pivot*.

Comecemos por trocar a primeira linha de A com a segunda, visto que $a_{21} > a_{11}$. Depois da primeira transformação, obtém-se a matriz $A^{(1)}$, da forma

$$A^{(1)} \;=\; \begin{bmatrix} 1 & 10^{-6} & 2 \\ 0 & -10^{-12} & 1 - 2 \times 10^{-6} \\ 0 & 2 - 10^{-6} & -3 \end{bmatrix}. \tag{3.49}$$

A pesquisa de pivot impõe que se troque a segunda linha com a terceira, visto que $a_{32} > a_{22}$. Depois de efectuar esta troca, realiza-se a segunda transformação da matriz, que nos leva ao sistema $A^{(2)}x = b^{(2)}$. Se os cálculos forem realizados com a precisão acima referida, resulta

$$\begin{aligned} A^{(2)} \;&=\; \begin{bmatrix} 1.00000 & 1.00000 \times 10^{-6} & 2.00000 \\ 0 & 2.00000 & -3.00000 \\ 0 & 0 & 9.99998 \times 10^{-1} \end{bmatrix}, \\ b^{(2)} \;&=\; \begin{bmatrix} 3.00000 \\ -1.00000 \\ 9.99997 \times 10^{-1} \end{bmatrix}. \end{aligned} \tag{3.50}$$

Resolvendo o sistema (3.50), obtém-se

$$\begin{aligned} x_3 \;&=\; \frac{9.99997 \times 10^{-1}}{1.00000} = 9.99999 \times 10^{-1} \\ x_2 \;&=\; \frac{-1.00000 + 3.00000\, x_3}{2.00000} = 1.00000 \\ x_1 \;&=\; 3.00000 \;-\; 2.00000\, x_3 \;-\; 1.00000 \times 10^{-6}\, x_2 = 9.99999 \times 10^{-1}. \end{aligned} \tag{3.51}$$

A solução agora calculada é bastante diferente da que obtivemos quando não foi utilizada a pesquisa de pivot. Se substituirmos estes valores no sistema (3.46), veremos que a nova solução está correcta, dentro dos limites da precisão utilizada.

Este exemplo mostra-nos como a pesquisa de pivot pode desempenhar um papel essencial no que respeita à minimização da instabilidade numérica quando se resolvem sistemas lineares pelo método da eliminação de Gauss. ♦

3.2.4 Métodos de factorização

Neste parágrafo vamos discutir alguns métodos directos que se baseiam na factorização da matriz dos coeficientes de um sistema linear $A\,x = b$.

Definição 3.6. Chama-se *factorização* LU de uma matriz não singular $A \in \mathbb{R}^{n\times n}$ à sua representação sob a forma do produto de duas matrizes,

$$A \;=\; L\,U,$$

onde L e U são matrizes triangulares, respectivamente inferior e superior.

Se for conhecida uma factorização LU de uma matriz A, o sistema linear $A\,x \;=\; b$ dá origem a dois sistemas lineares com matrizes dos coeficientes triangulares,

$$\begin{aligned} L\,g \;&=\; b \\ U\,x \;&=\; g, \end{aligned}$$

onde g é o vector auxiliar $g = U\,x$.

Além de nos permitir obter a a solução de sistemas lineares, a factorização LU tem outras aplicações, como por exemplo o cálculo de determinantes. Com efeito, o determinante de A é igual ao produto dos determinantes de L e de U, os quais se calculam imediatamente, já que estas matrizes são triangulares. De facto,

$$\begin{gathered} det\,L \;=\; l_{11}\,l_{22}\cdots l_{nn} \\ \text{e} \\ det\,U \;=\; u_{11}\,u_{22}\cdots u_{nn}, \end{gathered}$$

onde l_{ij} e u_{ij} designam respectivamente as entradas de L e de U.

Note-se que para calcularmos por definição o determinante de uma matriz de ordem n, teríamos de somar $n!$ parcelas, cada uma das quais é um produto de n entradas da matriz A. Tal cálculo significaria, por exemplo, que para uma matriz 10×10, deveríamos efectuar mais de 30 milhões de multiplicações! Compreende-se portanto que tal forma de cálculo de um determinante não seja aplicável na prática. Pelo contrário, se utilizarmos a referida factorização $L\,U$, o mesmo determinante pode ser calculado apenas com algumas centenas de operações aritméticas.

Uma vantagem suplementar dos métodos de factorização, uma vez factorizada uma matriz, consiste em podermos resolver vários sistemas diferentes com essa matriz, pois basta resolver os sistemas triangulares correspondentes (as matrizes L e U só precisam de ser determinadas uma vez). Isso é vantajoso, dado que, como vamos ver, nos métodos de factorização a determinação das matrizes L e U é precisamente a etapa mais dispendiosa, em termos de número de operações.

A factorização de uma matriz não singular $A \in \mathbb{R}^{n\times n}$ na forma LU não é única. Com efeito, podemos determinar L e U a partir de um sistema de n^2 equações,

$$a_{ij} \;=\; \sum_{k=1}^{n} l_{ik}\, u_{kj}, \quad i = 1:n, \quad j = 1:n, \tag{3.52}$$

onde l_{ik} e u_{kj} são incógnitas representando as componentes das matrizes L e U, respectivamente.

Uma vez que cada uma das matrizes L e U possui $\dfrac{n(n+1)}{2}$ entradas não nulas, o número total de incógnitas do sistema (3.52) é $n(n+1)$, portanto superior ao número de equações. O sistema (3.52) é por conseguinte indeterminado, isto é, admite uma infinidade de soluções. A cada uma dessas soluções corresponde uma certa factorização, que se caracteriza por um conjunto de condições suplementares.

Vamos analisar três casos particulares de factorização usados nas aplicações.

3.2.5 Factorização de Doolittle

Este tipo de factorização resulta de impormos as condições

$$l_{ii} \;=\; 1, \qquad i = 1:n. \tag{3.53}$$

Vamos mostrar como, a partir destas condições, se podem deduzir fórmulas para as entradas das matrizes L e U, as quais ficam assim completamente determinadas.

Seja a_{ij} uma qualquer entrada da matriz A, com $i \leq j$. Atendendo à forma triangular das matrizes L e U, bem como à condição (3.53), podemos escrever,

$$a_{ij} \;=\; \sum_{k=1}^{i} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj} \;+\; u_{ij}, \quad i = 1:n, \quad j = i:n. \tag{3.54}$$

Da fórmula (3.54), resulta imediatamente

$$u_{ij} \;=\; a_{ij} \;-\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj}\ . \tag{3.55}$$

A fim de deduzir uma fórmula análoga para a matriz L, consideremos uma qualquer entrada a_{ij}, com $i > j$. Neste caso, em vez de (3.54), temos

$$a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{j-1} l_{ik}\, u_{kj} \;+\; l_{ij}\, u_{jj}, \qquad i = 1:n, \quad j = i:n. \tag{3.56}$$

Donde, atendendo a que A é não singular (o mesmo acontecendo portanto com a matriz U), temos

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj}}{u_{jj}}. \tag{3.57}$$

Utilizando as expressões (3.55) e (3.57), podem calcular-se todas as entradas das matrizes L e U. Para isso, basta que todas as entradas da diagonal principal de U sejam diferentes de zero. Se, durante processo de cálculo, se obtiver alguma dessas entradas igual a zero, tal como acontece no método da eliminação de Gauss, deve-se proceder a alterações na matriz U. Neste caso podemos, por exemplo, alterar a ordem das colunas de U, mantendo a matriz L. Isto corresponde a trocar a ordem das colunas de A, ou seja, a trocar a ordem das incógnitas do sistema $A\,x = b$.

Ao calcular o determinante de A com base numa factorização $L\,U$, deve-se entrar em conta com as permutações efectuadas das linhas ou colunas. Assim,

$$\det A = (-1)^{Nt} \det L \times \det U, \tag{3.58}$$

onde Nt é o número de trocas de colunas efectuadas.

A troca de colunas de L também pode ser aplicada para atenuar os problemas de instabilidade numérica que podem ocorrer durante o cálculo dos factores L e U. Para esse efeito pode usar-se a mesma estratégia da pesquisa parcial de pivot atrás descrita.

É interessante notar que o método da eliminação de Gauss é algebricamente equivalente ao método de Doolittle[12], podendo, neste sentido, ser considerado também um método de factorização. Para verificarmos isso, recordemos que no método da eliminação de Gauss se obtém uma matriz triangular superior U, dada pela fórmula (3.33), pág. 106. Além disso, durante o cálculo da matriz U são utilizados os coeficientes m_{ik}, para $k = 1 : n$ e $i = k + 1 : n$, definidos pela fórmula (3.32).

Se construirmos uma matriz triangular inferior cujas entradas na diagonal principal são todas iguais a 1, e as restantes entradas sejam os coeficientes m_{ij}, obtemos a seguinte matriz L,

$$L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \dots & m_{n-1,n-2} & 1 & 0 \\ \dots & m_{n,n-2} & m_{n,n-1} & 1 \end{bmatrix}. \tag{3.59}$$

A discussão acima leva-nos ao seguinte

[12]Myrick Hascall Doolittle, 1830-1911, matemático americano.

Teorema 3.3. As matrizes L e U, dadas respectivamente pelas fórmulas (3.59) e (3.33), pág. 106, produzem a factorização $A = LU$, idêntica à factorização de Doolittle.

Demonstração. Vamos demonstrar as igualdades

$$a_{ij}^{(i-1)} \,=\, u_{ij}, \qquad i = 1:n, \quad j = i:n \tag{3.60}$$

e

$$m_{ij} \,=\, l_{ij}, \quad j = 1:n, \quad i = j:n. \tag{3.61}$$

Para isso, basta comparar as fórmulas do método de Gauss com as da factorização de Doolittle. Em primeiro lugar, sabemos que

$$a_{1j} \,=\, u_{1j}, \quad j = 1:n, \qquad m_{i1} \,=\, \frac{a_{i1}}{a_{11}}, \quad i = 2:n.$$

Usando indução, vamos supor que as igualdades (3.60) são satisfeitas para as linhas da matriz U, com índice $k = 1, \ldots, i-1$, e que as igualdades (3.61) se verificam para todas as colunas de L, com índice $k = 1, \ldots, j-1$.

Verifiquemos que as mesmas identidades se mantêm válidas para a i-ésima linha de U e para a j-ésima coluna de L. De facto, de acordo com a fórmula (3.31), pág. 105, do método de Gauss, temos

$$a_{ij}^{(k)} \,=\, a_{ij}^{(k-1)} \,-\, m_{ik}\, a_{kj}^{(k-1)}, \quad k = 1:n-1, \tag{3.62}$$

onde se subentende que $a_{ij}^{(0)} \,=\, a_{ij}$, para $i = 1:n$ e $j = 1:n$. Aplicando a fórmula (3.62) sucessivamente, com $k = 1, \ldots, i-1$, obtém-se

$$\begin{array}{ll} a_{ij}^{(1)} & = a_{ij} \,-\, m_{i1}\, a_{1j} \\ a_{ij}^{(2)} & = a_{ij}^{(1)} \,-\, m_{i2}\, a_{2j}^{(1)} a_{ij} \,-\, m_{i1}\, a_{1j} \,-\, m_{i2}\, a_{2j}^{(1)} \\ & \vdots \\ a_{ij}^{(i-1)} & = a_{ij}^{(i-2)} \,-\, m_{i,i-1}\, a_{i-1,j}^{(i-2)} a_{ij} \,-\, \sum_{k=1}^{i-1} m_{ik}\, a_{kj}^{(k-1)}. \end{array} \tag{3.63}$$

Se, de acordo com a hipótese de indução, substituirmos os coeficientes $m_{i,k}$ e $a_{k,j}^{(k-1)}$, no segundo membro de (3.63), por l_{ik} e u_{kj}, obtemos a fórmula (3.55), donde se conclui que $a_{ij}^{(i-1)} \,= u_{ij}$, com $j = i, \ldots, n$.

Considerando agora as entradas da j-ésima coluna de L, de acordo com (3.32), pág. 106, elas têm a forma

$$m_{ij} \,=\, \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}, \quad i = j:n. \tag{3.64}$$

Analogamente à dedução da fórmula (3.63), podemos mostrar que

$$a_{ij}^{(j-1)} \;=\; a_{ij} \;-\; \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}. \tag{3.65}$$

Se, no segundo membro de (3.64), substituirmos o numerador de acordo com (3.65), obtemos

$$m_{ij} \;=\; \frac{a_{ij} \;-\; \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}}{a_{jj}^{(j-1)}}, \quad i = j : n. \tag{3.66}$$

Mas, atendendo à hipótese de indução, podemos substituir no segundo membro de (3.66), $a_{kj}^{(k-1)}$ por u_{kj}, para $k = 1 : j$, e m_{ik} por l_{ik}, para $k = 1 : i$. Então, o segundo membro de (3.66) fica igual ao segundo membro de (3.57), de onde se conclui que $m_{ij} \;= l_{ij}$, para todas as componentes da j-ésima coluna da matriz L. Fica assim provada, por indução, a afirmação do teorema. □

Do Teorema 3.3 resulta que os métodos de Gauss e de Doolittle são idênticos, no sentido em que na resolução de um sistema linear segundo cada um desses métodos, efectuam-se exactamente as mesmas operações aritméticas. Em particular, para o sistema $A\,x = b$, as três etapas que distinguimos no método de Gauss coincidem com as etapas do método de Doolittle (ou de qualquer outro método de factorização), a saber:

1. Factorização $L\,U$ da matriz A;
2. Resolução do sistema $L\,g \;= b$;
3. Resolução do sistema $U\,x \;= g$.

Por conseguinte, de acordo com o que dissemos em relação ao método de Gauss, podemos concluir que a etapa mais dispendiosa dos cálculos, quando se aplica o método de Doolittle, é a primeira – exigindo cerca de $2n^3/3$ operações aritméticas. As outras duas etapas requerem cerca de n^2 operações cada uma. As mesmas conclusões são aplicáveis à factorização de Crout, de que nos ocupamos a seguir.

3.2.6 Factorização de Crout

Outro tipo comum de factorização, a chamada *factorização de Crout*[13], baseia-se na imposição das seguintes condições sobre a diagonal principal da matriz U:

$$u_{ii} = 1, \quad i = 1 : n.$$

[13]Prescott Durand Crout, 1907 -1984, matemático americano.

As fórmulas para as entradas das matrizes L e U da factorização de Crout deduzem-se da mesma maneira que no caso da factorização de Doolittle. Assim, no caso de $i \geq j$, são válidas as igualdades

$$a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{j-1} l_{ik}\, u_{kj} \;+\; l_{ij}, \qquad j = 1:n, \quad i = j:n.$$

Daqui obtém-se imediatamente

$$l_{ij} \;=\; a_{ij} \;-\; \sum_{k=1}^{j-1} l_{ik}\, u_{kj}\ . \tag{3.67}$$

No que diz respeito à matriz L, partimos da igualdade

$$a_{ij} \;=\; \sum_{k=1}^{i} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj} \;+\; l_{ii}\, u_{ij}, \qquad i = 1:n, \quad j = 1:i. \tag{3.68}$$

Da igualdade (3.68) resulta

$$u_{ij} \;=\; \frac{a_{ij} \;-\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj}}{l_{ii}}\ . \tag{3.69}$$

As fórmulas (3.67) e (3.69), quando aplicadas alternadamente (começando com a primeira coluna de L e acabando com a última linha de U), permitem-nos determinar completamente as matrizes L e U da factorização de Crout, desde que se verifique $l_{ii} \neq 0$, para $i = 1:n$.

Se durante o processo de factorização acontecer que $l_{ii} = 0$, para um certo i, procede-se a uma troca de linhas na matriz L, mantendo U sem alteração. Esta troca é acompanhada pela mesma permutação das linhas da matriz A e das entradas do segundo membro do sistema. Tal como no caso da factorização de Doolittle, tais permutações implicam uma troca de sinal no cálculo do determinante, de acordo com (3.58), pág. 116.

Também no caso da factorização de Crout é conveniente aplicar a pesquisa parcial de pivot, efectuando-se trocas de linhas quando os elementos diagonais l_{ii} forem pequenos em módulo.

Exemplo 3.5. *Dado o sistema $A\,x = b$, onde*

$$A \;=\; \begin{bmatrix} 2 & 1 & 3 \\ -2 & -1 & 1 \\ 2 & 4 & 2 \end{bmatrix}, \qquad b \;=\; \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix},$$

pretende-se determinar a solução mediante aplicação de factorização de Doolittle e de Crout.

Comecemos por factorizar A segundo o método de Doolittle. Tal como resulta da fórmula (3.55), pág. 115, a primeira linha de U é igual à primeira linha de A, ou seja,

$$u_{11} = 2, \qquad u_{12} = 1, \qquad u_{13} = 3 \ .$$

Calculando os elementos da primeira coluna de L, de acordo com a fórmula (3.57), obtemos

$$l_{11} = 1, \quad l_{21} = \frac{a_{21}}{u_{11}} = -1, \quad l_{31} = \frac{a_{31}}{u_{11}} = 1 \ .$$

Passemos ao cálculo da segunda linha de U. Temos

$$\begin{array}{ll} u_{22} & = a_{22} - l_{21}\, u_{12} = 0 \\ u_{23} & = a_{23} - l_{21}\, u_{13} = 4 \ . \end{array}$$

Como sabemos, sendo $u_{22} = 0$, não é possível prosseguir os cálculos sem alterar a matriz A. Assim, uma vez que $u_{23} \neq 0$, vamos trocar de lugar a segunda com a terceira coluna de U, fazendo simultaneamente a mesma troca em A. Sejam U' e A', respectivamente, as matrizes resultantes. Podemos escrever

$$\begin{array}{rcl} u'_{22} & = & u_{23}, \\ u'_{23} & = & u_{22} \ . \end{array}$$

Continuando o processo de factorização com as matrizes U' e A', obtém-se

$$\begin{array}{ll} l_{32} & = \dfrac{a'_{32} - l_{31}\, u'_{12}}{u'_{22}} = \dfrac{a_{33} - l_{31}\, u_{13}}{u_{23}} = -\dfrac{1}{4} \\ u'_{33} & = a'_{33} - l_{31}\, u'_{13} - l_{32}\, u'_{23} = a_{32} - l_{31}\, u_{12} - l_{32}\, u_{22} = 3 \ . \end{array}$$

Recapitulando, obtivemos a seguinte factorização de A:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -\frac{1}{4} & 1 \end{bmatrix}, \quad U' = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Para calcular a solução do sistema dado, comecemos por resolver o sistema com a matriz triangular inferior $L\, g = b$, de acordo com o método habitual.

$$\begin{array}{rcl} g_1 & = b_1 \iff & g_1 = 5 \\ -g_1 + g_2 & = b_2 \iff & g_2 = 4 \\ g_1 - g_2/4 + g_3 & = b_3 \iff & g_3 = 0 \ . \end{array}$$

Ao resolver o sistema $U' x = g$, temos de ter em conta que a segunda coluna de U trocou de lugar com a terceira. Isto equivale a uma troca de posições entre x_2 e x_3. Assim, temos

$$\left\{ \begin{array}{llll} 2\, x_1 & +3\, x_3 & +x_2 & = g_1 \\ & 4x_3 & & = g_2 \\ & & 2\, x_3 & = g_3, \end{array} \right.$$

donde $x_2 = 0$, $x_3 = 1$ e $x_1 = 1$. Se em vez do método de Doolittle quisermos aplicar a *factorização de Crout*, teremos de basear os cálculos nas fórmulas (3.67) e (3.69), pág 119. Nesse caso, a primeira coluna de L fica igual à primeira coluna de A.

Para a primeira linha de U, obtém-se

$$u_{11} = 1, \qquad u_{12} = \frac{a_{12}}{l_{11}} = \frac{1}{2}, \qquad u_{13} = \frac{a_{13}}{l_{11}} = \frac{3}{2} \; .$$

Na segunda coluna de L, tém-se

$$\begin{array}{rcrcl} l_{22} & = & a_{22} - l_{21}\, u_{12} & = & 0 \\ l_{32} & = & a_{32} - l_{31}\, u_{12} & = & 3 \; . \end{array}$$

Uma vez que $l_{22} = 0$, torna-se necessário trocar a segunda com a terceira linha de L (e, consequentemente, de A). Obtemos

$$\begin{array}{rl} l'_{22} & = l_{32} = 3 \\ l'_{32} & = l_{22} = 0 \; . \end{array}$$

Resta calcular as componentes da segunda linha de U e terceira coluna de L,

$$u_{23} = \frac{a'_{23} - l'_{21} u_{13}}{l'_{22}} = -\frac{1}{3}$$

$$l'_{33} = a'_{33} - l'_{31} u_{13} - l'_{32} u_{23} = 4 \; .$$

Consequentemente, a factorização de Crout da matriz dada tem a forma

$$L' = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ -2 & 0 & 4 \end{bmatrix}, \qquad U = \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix} .$$

A partir de qualquer uma das factorizações de A obtidas, utilizando a fórmula (3.58), pág. 116, calcula-se facilmente o determinante de A,

$$det\, A = det\, L' \, (-1)^1 = det\, U' \, (-1)^1 = -24 \; .$$

Para resolver o sistema dado com base na factorização de Crout, basta considerar o segundo membro $b' = (5, 4, -1)^T$ (uma vez que foi trocada a segunda com a terceira linha de U), após o que se resolvem os sistemas $L'\, g = b'$ e $U\, x = g$, utilizando substituições descendentes (para o primeiro sistema) e substituições ascendentes (para o segundo). ♦

3.2.7 Factorização de Cholesky

Os dois tipos de factorização que referimos anteriormente existem para qualquer matriz não singular (ainda que possa ser necessário efectuar uma troca de linhas ou colunas).

Quanto à *factorização de Cholesky*[14], que vamos discutir a seguir, só é aplicável a matrizes (simétricas) *definidas positivas*[15]. Embora se trate de uma restrição muito forte, este tipo de factorização não deixa de ter interesse prático, visto que tais matrizes ocorrem em muitos problemas de cálculo numérico, por exemplo, no *método dos mínimos quadrados* e em certos *problemas de valores de fronteira* para equações diferenciais.

A maior vantagem deste tipo de factorização consiste em só necessitarmos de calcular uma matriz triangular L, visto que uma matriz simétrica definida positiva pode ser representada sob a forma $A = L\,L^T$. Isto significa que o número de operações para resolver um sistema linear fica reduzido a cerca de metade, quando se compara o método de Cholesky com outros métodos de factorização, ou com o método de Gauss.

A factorização de Cholesky baseia-se no teorema a seguir.

Teorema 3.4. Seja $A \in \mathbb{R}^{n\times n}$ uma matriz simétrica definida positiva. Então a matriz A pode ser factorizada na forma

$$A = \tilde{L}\,\tilde{D}\,\tilde{L}^T, \tag{3.70}$$

onde $\tilde{L}$ é uma matriz triangular inferior com 1's na diagonal e $\tilde{D}$ é uma matriz diagonal com todas as entradas diagonais positivas.

A matriz A pode também escrever-se na forma

$$A = L\,L^T, \tag{3.71}$$

onde L é uma matriz triangular inferior.

Demonstração. Uma vez provado (3.70), a factorização (3.71) é imediata já que basta tomar em (3.70) $L = \tilde{L}\,\tilde{D}^{1/2}$. Esta matriz L está bem definida porquanto as entradas na diagonal principal de $\tilde{D}$ são positivas.

Provemos agora a existência da factorização (3.70). A prova será realizada por indução sobre a ordem k da matriz A. Para $k = 1$, a igualdade (3.70) é trivialmente satisfeita, visto que

$$A = [a_{11}] = \underbrace{[1]}_{\tilde{L}}\ \underbrace{[a_{11}]}_{\tilde{D}}\ \underbrace{[1]}_{\tilde{L}^T}.$$

[14] André - Louis Cholesky, 1875-1918, militar e matemático francês.

[15] Sobre esta classe fundamental de matrizes, ver adiante o parágrafo 3.6, pág. 163.

Suponhamos que para $k = n-1$ se verifica (3.70), isto é,

$$A_{n-1} = \tilde{L}_{n-1}\, \tilde{D}_{n-1}\, \tilde{L}_{n-1}^T, \tag{3.72}$$

onde A_{n-1} é uma matriz simétrica definida positiva, de ordem $n-1$, e as matrizes $\tilde{L}_{n-1}$ e $\tilde{D}_{n-1}$ verificam as condições de $\tilde{L}$ e $\tilde{D}$ no enunciado.

A matriz A, de ordem n, pode escrever-se na forma

$$A = \begin{bmatrix} A_{n-1} & c \\ c^T & a_{nn} \end{bmatrix}, \qquad \text{onde} \quad c \in \mathbb{R}^{n-1},$$

e A_{n-1} é a submatriz que resulta de A suprimindo a última linha e a última coluna. Como A é definida positiva, A_{n-1} também o é (e portanto admite a factorização (3.72)). Considere-se a igualdade

$$A = \underbrace{\begin{bmatrix} \tilde{L}_{n-1} & 0 \\ c^T\, \tilde{L}_{n-1}^{-T}\, \tilde{D}_{n-1}^{-1} & 1 \end{bmatrix}}_{\tilde{L}} \underbrace{\begin{bmatrix} \tilde{D}_{n-1} & 0 \\ 0 & \alpha \end{bmatrix}}_{\tilde{D}} \underbrace{\begin{bmatrix} \tilde{L}_{n-1}^T & \tilde{D}_{n-1}^{-1}\, \tilde{L}_{n-1}^{-1}\, c \\ 0 & 1 \end{bmatrix}}_{\tilde{L}^T}. \tag{3.73}$$

A matriz $\tilde{L}$ tem a forma pretendida (triangular inferior com 1's na diagonal) e está bem definida já que as matrizes $\tilde{L}_{n-1}$ e $\tilde{D}_{n-1}$ são obviamente invertíveis.
Resta provar que a entrada α em $\tilde{D}$ é positiva, para se concluir que a matriz $\tilde{D}$ possui as entradas diagonais positivas.
Uma vez que a matriz A é definida positiva, conclui-se de (3.73) ser válida a desigualdade

$$0 < \det(A) = \det(\tilde{L}_{n-1})\, \det(\tilde{D})\, \det(\tilde{L}^T) = 1 \times \det(\tilde{D}) \times 1.$$

Atendendo a que $\det(\tilde{D}) = \alpha\, \det(\tilde{D}_{n-1})$ e, por hipótese de indução, $\det(\tilde{D}_{n-1}) > 0$, resulta que $\alpha > 0$. Por conseguinte, a matriz A pode factorizar-se na forma (3.70). □

Observação

Note-se que em resultado da demonstração anterior, a matriz L da factorização (3.71) pode ser escolhida por forma que as entradas da sua diagonal principal sejam positivas. No entanto, se partirmos de uma factorização como

$$A = \underbrace{\begin{bmatrix} \hat{L} & 0 \\ \gamma^T & z \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} \hat{L}^T & \gamma \\ 0 & z \end{bmatrix}}_{L^T},$$

onde $\hat{L}$ é uma matriz triangular inferior, de ordem $n-1$, e com determinante positivo, tem-se

$$\det(A) = z^2\, (\det \hat{L})^2 \Longrightarrow z = \pm \frac{\sqrt{\det(A)}}{\det(\hat{L})}.$$

Neste caso, escolhe-se a raiz positiva de modo que todos os elementos da diagonal principal de L são positivos.

Fórmulas computacionais para a factorização de Cholesky

Vejamos, em termos práticos, como se pode calcular a matriz L da factorização de Cholesky. Seja a_{ij} uma entrada de A, com $i \geq j$. Da igualdade (3.71) resulta

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, l_{jk} = \sum_{k=1}^{j-1} l_{ik}\, l_{jk} + l_{ij}\, l_{jj}, \qquad j = 1 : n, \qquad i = j : n. \tag{3.74}$$

No caso de $i = j$, da igualdade (3.74) obtém-se a fórmula para as entradas da diagonal principal de L,

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}, \qquad i = 1 : n. \tag{3.75}$$

De acordo com o Teorema 3.2, pág. 100, todos os elementos da diagonal principal de L são reais, pelo que o segundo membro de (3.75) é sempre real.

Uma vez calculado l_{jj}, podemos obter as restantes entradas da j-ésima coluna de L. Da fórmula (3.74) obtém-se,

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, l_{jk}}{l_{jj}}, \qquad i = j + 1 : n. \tag{3.76}$$

Assim, usando as fórmulas (3.75) e (3.76) alternadamente, pode ser obtida a factorização de Cholesky da matriz A.

Exemplo 3.6. *Consideremos a matriz de ordem n,*

$$A = \begin{bmatrix} 4 & 2 & 0 & \dots & 0 \\ 2 & 5 & 2 & \dots & 0 \\ 0 & 2 & 5 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 2 & 5 & 2 \\ 0 & \dots & 0 & 2 & 5 \end{bmatrix}.$$

Trata-se de uma matriz simétrica tridiagonal*, isto é*

$$a_{ij} \neq 0 \Rightarrow |i - j| \leq 1.$$

Matrizes com estas características aparecem frequentemente nas aplicações. Vamos obter a sua factorização de Cholesky.

Dado não ser imediato decidir se a matriz dada é definida positiva, vamos tentar utilizar as fórmulas (3.75) e (3.76) e verificar se elas são aplicáveis. No caso afirmativo poderemos estar certos da positividade da matriz A.

Comecemos pela entrada l_{11}. De acordo com (3.75), o seu valor é

$$l_{11} \,=\, \sqrt{a_{11}} \,=\, 2 \;. \tag{3.77}$$

As restantes entradas da primeira coluna são dadas pela fórmula (3.76),

$$\begin{aligned} l_{21} \;&= \frac{a_{21}}{l_{11}} = 1 \\ \\ l_{k1} \;&= \frac{a_{k1}}{l_{11}} = 0, \qquad k = 3 : n. \end{aligned}$$

Vamos provar por indução que as restantes colunas da matriz L têm a mesma estrutura, isto é, para a coluna j verifica-se,

$$\begin{array}{ll} l_{jj} & = 2 \\ l_{j+1,j} & = 1 \\ l_{i,j} & = 0, \qquad i = j+2 : n. \end{array} \tag{3.78}$$

Para a primeira coluna, as fórmulas (3.78) já estão provadas. Suponhamos agora que estas fórmulas são válidas para todas as colunas, até à de ordem $j-1$.

Vejamos o que acontece com a coluna j. De acordo com a fórmula (3.75), podemos escrever

$$l_{jj} \,=\, \sqrt{a_{jj} \,-\, \sum_{k=1}^{j-1} l_{jk}^2} \,=\, \sqrt{a_{jj} \,-\, l_{j,j-1}^2} \,=\, 2 \;.$$

Aplicando a fórmula (3.76), obtemos

$$\begin{array}{ll} l_{j+1,j} & = \dfrac{a_{j+1,j}}{l_{jj}} = 1 \\ l_{i,j} & = 0, \qquad i = j+2, \ldots, n. \end{array}$$

Fica assim provado que a factorização de Cholesky da matriz dada é definida por uma matriz triangular inferior com a forma

$$L = \begin{bmatrix} 2 & 0 & 0 & \ldots & 0 \\ 1 & 2 & 0 & \ldots & 0 \\ 0 & 1 & 2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & 1 & 2 & 0 \\ 0 & \ldots & 0 & 1 & 2 \end{bmatrix}.$$

O determinante de A pode ser calculado com base nessa factorização, obtendo-se

$$det\, A \,=\, (det\, L)^2 \,=\, (l_{11}\, l_{22}\, \ldots\, l_{nn})^2 \,=\, (2^n)^2 \,=\, 4^n. \tag{3.79}$$

Uma vez que a fórmula (3.79) é válida para qualquer n, ela pode servir para calcularmos os menores principais da matriz A dada. Assim, temos

$$A_1 = 4,\ A_2 = 4^2,\ \ldots, A_n = det\, A = 4^n.$$

Fica assim provado que todos os menores principais de A são positivos, de onde resulta que A é definida positiva (ver Teorema 3.13, pág. 165). ⧫

3.3 Métodos iterativos para sistemas lineares

Nesta secção vamos estudar alguns métodos iterativos para o cálculo aproximado de soluções de sistemas lineares. Começamos por apresentar alguns conceitos gerais que serão úteis posteriormente.

3.3.1 Noções básicas sobre métodos iterativos

Em certos problemas matemáticos e nas aplicações, quando se revela impossível ou muito difícil calcular a solução exacta de um problema, opta-se por se tentar obter um valor aproximado dessa solução. Esse valor aproximado é geralmente calculado mediante um *método de aproximações sucessivas*, ou *método iterativo*, onde cada nova aproximação é obtida a partir da anterior (ou das anteriores). Pretende-se deste modo tornar o erro de cada aproximação tão pequeno quanto se queira.

A definição a seguir caracteriza o conceito de método iterativo num espaço normado.

Definição 3.7. Seja E um espaço normado e X um subconjunto de E. Chama-se *método iterativo de p passos* em E, uma aplicação Ψ que a cada vector de p componentes, $(\xi_0, \ldots, \xi_{p-1}) \in X$, faz corresponder uma sucessão $(x^{(k)})_{k\geq k_0}$, onde $x^{(k)} \in E$, com as seguintes propriedades:

1. Os primeiros p termos são os dados,

$$x^{(i)} = \xi_i, \qquad i = 0, \ldots, p-1.$$

2. Os restantes elementos da sucessão $(x^{(k)})_{k\geq k_0}$ são obtidos a partir dos dados, de acordo com a fórmula

$$x^{(k+p)} = \phi(x^k, x^{k+1}, \ldots, x^{k+p-1}),$$

onde ϕ é uma função dada (chamada *função iteradora*), com domínio em X e valores em E.

Estamos fundamentalmente interessados em métodos iterativos definidos em $E = \mathbb{R}^n$ munido das normas usuais, e de um passo. Na prática apenas se calcula um número finito de termos da sucessão $\left(x^{(k)}\right)_{k\geq k_0}$ (também chamados *iteradas*), tantos quantos necessários para alcançar a precisão pretendida. Por isso, a cada método iterativo estão geralmente associados *critérios de paragem*, isto é, regras que nos permitem verificar se uma dada iterada possui ou não a precisão exigida.

Convergência

O conceito de convergência de um método iterativo é fundamental.

Definição 3.8. Dizemos que um método iterativo de p passos, definido sobre $X \subseteq \mathbb{R}^n$, é convergente para um certo $x \in \mathbb{R}^n$, se para quaisquer valores iniciais $(\xi_0, \ldots, \xi_{p-1})$, se verificar $x^{(k)} \to x$, quando $k \to \infty$ (segundo a norma adoptada em $\mathbb{R}^n$), isto é, $\lim_{k\to\infty} ||x - x^{(k)}|| = 0$.

Sabe-se que a convergência em espaços de dimensão finita não depende da norma considerada (ver prova por exemplo em [28], p. 8). Daí que, no caso dos métodos iterativos para sistemas lineares, que vamos estudar nos próximos parágrafos, a convergência numa certa norma é equivalente à convergência noutra norma qualquer que adoptemos.

Resulta da Definição 3.8 que o método iterativo *não* converge desde que exista pelo menos um elemento inical x_0, para o qual a sucessão $(x_k)_{k\geq 0}$ não é convergente.

Estabilidade

Além da convergência, outra propriedade importante dos métodos iterativos é a sua *estabilidade*. Um método iterativo que parta de dois vectores iniciais ξ e η, que sejam "próximos", se as respectivas iteradas do método se mantêm próximas, diz-se um método *estável*, no sentido da definição a seguir.

Por exemplo, um processo iterativo que na passagem de um vector inicial x_0 ao vector $fl(x_0)$, conduza a vectores de iteradas que não sejam respectivamente próximas das que se obteriam caso não houvesse lugar a arredondamentos, deverá ser considerado instável.

Definição 3.9. Um método iterativo Ψ, de p passos, definido no conjunto X, diz-se *estável em* $B \subset X$, se existir uma constante $c > 0$, tal que

$$\max_{n\in N} \|x^{(n)} - y^{(n)}\| \leq c \max_{i=1,\ldots,p} \|\xi_i - \eta_i\| \qquad \forall\, \xi, \eta \in B, \tag{3.80}$$

onde $(x_n)_{n\geq 0}$ e $(y_n)_{n\geq 0}$ são, respectivamente, as sucessões geradas a partir de $\xi = (\xi_0, \xi_1, \ldots, \xi_{p-1})$ e $\eta = (\eta_0, \eta_1, \ldots, \eta_{p-1})$.

Para representar o erro da k-ésima iterada usaremos a notação $e^{(k)}$, ou seja, $e^{(k)} = x - x^{(k)}$.

3.3.2 Métodos iterativos para sistemas lineares

Nos próximos parágrafos vamos analisar alguns métodos iterativos para o cálculo aproximado da solução do sistema linear

$$A\,x \;=\; b, \tag{3.81}$$

onde $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^{n\times 1}$.

Supomos que a matriz A é não singular, pelo que o sistema (3.81) tem uma única solução.

Com o objectivo de construir um método iterativo, começamos por reduzir o sistema (3.81) a uma forma equivalente

$$x = G(x) = Cx + g, \tag{3.82}$$

onde C é uma certa matriz (a que chamaremos *matriz de iteração*), e g é um vector auxiliar ($g \in \mathbb{R}^{n\times 1}$).

Uma vez escrito o sistema na forma (3.82), podemos dizer que a sua solução é um ponto fixo da função G (definida em $\mathbb{R}^n$ e com valores no mesmo espaço). A ideia é determinar o ponto fixo de G por um método análogo ao método do ponto fixo, utilizado no capítulo anterior para aproximar os pontos fixos de funções de uma variável.
Assim, dada uma certa aproximação inicial $x^{(0)}$, vamos construir uma sucessão de vectores através da fórmula de recorrência,

$$x^{(k+1)} = G(x^{(k)}) = Cx^{(k)} + g, \qquad k = 0, 1, \ldots \tag{3.83}$$

Tal transformação do sistema pode ser feita de muitas maneiras dando consequentemente origem a diferentes métodos iterativos, os quais podem ou não convergir.

O Teorema do ponto fixo em $\mathbb{R}^n$ será discutido mais tarde (ver pág. 172). Vamos no entanto antecipar desde já esse resultado fundamental, porquanto ele encontra uma aplicação natural nos processos iterativos do tipo (3.83) para aproximação da solução de um sistema linear.
Com efeito, o espaço linear $D = \mathbb{R}^n$ é fechado e convexo[16] (o que generaliza a noção de intervalo $I = [a, b] \subset \mathbb{R}$), e a função G em (3.83) aplica um vector $x \in D$

[16]Um conjunto X diz-se convexo se, para quaisquer x_1, x_2 pertencentes a X, todos os pontos do segmento $[x_1, x_2]$ também pertencerem a X. Isto é, o ponto $w = x_1 + t\,(x_2 - x_1)$, com $0 \le t \le 1$, pertence a X sempre que x_1 e x_2 pertencem a X.

num vector $y = G(x) \in D$, ou seja, $G(D) \subset D$. Além disso, a função linear (3.82) é de classe C^1 em D, e

$$G'(x) = \left[\frac{\partial G_i}{\partial x_j}\right]_{i,j=1}^{n}(x) = C, \qquad \forall x \in \mathbb{R}^n.$$

Assim, uma vez fixada uma norma vectorial e a correspondente norma matricial induzida, tem-se

$$||G'(x)|| = ||C||, \qquad \forall x \in \mathbb{R}^n.$$

A igualdade anterior não depende do ponto x considerado. Consequentemente, aplicando o Teorema do ponto fixo em $\mathbb{R}^n$, podemos afirmar que, na hipótese da matriz de iteração C ser tal que

$$||C|| < 1,$$

a equação (3.82) tem uma única solução e o processo iterativo $x^{(k+1)} = G\left(x^{(k)}\right)$ converge para essa solução, independentemente da escolha que se fizer da aproximação inicial $x^{(0)}$.[17] São válidas as seguintes majorações de erro:

$$||C|| < 1 \Longrightarrow \begin{cases} 1. & ||x - x^{(k+1)}|| \leq ||C||\,||x - x^{(k)}|| \\ 2. & ||x - x^{(k)}|| \leq ||C||^k\,||x - x^{(0)}|| \\ 3. & ||x - x^{(k+1)}|| \leq \dfrac{||C||}{1 - ||C||}\,||x^{(k+1)} - x^{(k)}|| \\ 4. & ||x - x^{(k)}|| \leq \dfrac{||C||^k}{1 - ||C||}\,||x^{(1)} - x^{(0)}|| \end{cases} \tag{3.84}$$

Descrevemos a seguir três métodos do tipo (3.83).

[17] Note que $||C||$ tem neste contexto um papel análogo ao da constante L usada no Teorema do ponto fixo em $\mathbb{R}$, pág. 47.

3.3.3 Método de Jacobi

Para deduzirmos as fórmulas iterativas do método de Jacobi[18], começamos por reescrever o sistema (3.81), pág. 128, na forma

$$\begin{aligned} x_1 &= \frac{b_1 - a_{12}\,x_2 - a_{13}\,x_3 - \cdots - a_{1n}\,x_n}{a_{11}} \\ x_2 &= \frac{b_2 - a_{21}\,x_1 - a_{23}\,x_3 - \cdots - a_{2n}\,x_n}{a_{22}} \\ &\vdots \\ x_n &= \frac{b_n - a_{n1}\,x_1 - a_{n2}\,x_2 - \cdots - a_{n,n-1}\,x_{n-1}}{a_{nn}}\ . \end{aligned} \tag{3.85}$$

O sistema (3.85) é equivalente ao inicial e é da forma $x = G(x)$. Note que assumimos serem não nulos todos os elementos da diagonal principal da matriz A, isto é, $a_{ii} \neq 0$, para $i = 1 : n$.
Se considerarmos a função iteradora G correspondente ao sistema (3.85), obtêm-se as seguintes *fórmulas computacionais*:

$$\begin{aligned} x_1^{(k+1)} &= \frac{b_1 - a_{12}\,x_2^{(k)} - a_{13}\,x_3^{(k)} - \cdots - a_{1n}\,x_n^{(k)}}{a_{11}} \\ x_2^{(k+1)} &= \frac{b_2 - a_{21}\,x_1^{(k)} - a_{23}\,x_3^{(k)} - \cdots - a_{2n}\,x_n^{(k)}}{a_{22}} \qquad , \qquad k = 0, 1, 2, \ldots \\ &\vdots \\ x_n^{(k+1)} &= \frac{b_n - a_{n1}\,x_1^{(k)} - a_{n2}\,x_2^{(k)} - \cdots - a_{n,n-1}\,x_{n-1}^{(k)}}{a_{nn}}\ . \end{aligned} \tag{3.86}$$

As expressões (3.86) podem escrever-se na seguinte forma compacta,

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} - \frac{\sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)}}{a_{ii}}, \qquad i = 1 : n, \quad k = 0, 1, 2, \ldots \tag{3.87}$$

Assim, o processo pode escrever-se matricialmente na forma $x^{(k+1)} = C_J\,x^{(k)} + g_J$, onde

$$C_J = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \vdots & \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{bmatrix} \quad \text{e} \quad g_J = \begin{bmatrix} b_1/a_{11} \\ b_2/a_{22} \\ \vdots \\ b_n/a_{nn} \end{bmatrix}.$$

[18] Carl Gustav Jacob Jacobi, 1804-1851, matemático alemão.

A formulação matricial deste e de outros processos iterativos será retomada adiante. O método de Jacobi, sendo o mais simples, permite a escrita imediata da respectiva matriz C_J e do vector constante g_J, directamente a partir das expressões (3.86).

Sublinhe-se desde já que no método de Jacobi a diagonal principal da respectiva matriz de iteração C_J possui entradas nulas, e que fora da diagonal se encontram os simétricos da matriz A do sistema, divididos pelo "pivot" da respectiva linha.

Exemplo 3.7. *Consideremos o sistema $A\,x = b$, onde*

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \quad e \quad b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}. \tag{3.88}$$

(a) Efectuar uma iteração do método de Jacobi, tomando como aproximação inicial $x^{(0)} = (0.5, 0.8, 1)$.

(b) Sabendo que a solução exacta do sistema é $x = (0.583, 0.833, 0.917)$, calcular

$$\|e^{(0)}\|_1 \qquad e \qquad \|e^{(1)}\|_1.$$

(a) Do sistema dado resultam as seguintes fórmulas iterativas,

$$\begin{aligned} x_1^{(1)} &= \frac{b_1 - a_{12}\,x_2^{(0)} - a_{13}\,x_3^{(0)}}{a_{11}} = \frac{1}{2}\,(2 - 0.8 - 0) = 0.6 \\ x_2^{(1)} &= \frac{b_2 - a_{21}\,x_1^{(0)} - a_{23}\,x_3^{(0)}}{a_{22}} = \frac{1}{2}(2 + 0.5 - 1) = 0.75 \\ x_3^{(1)} &= \frac{b_3 - a_{31}\,x_1^{(0)} - a_{32}\,x_2^{(0)}}{a_{33}} = \frac{1}{2}(1 - 0 + 0.8) = 0.9\ . \end{aligned}$$

A matriz de iteração obtém-se imediatamente a partir das fórmulas computacionais do método:

$$C_J = \begin{bmatrix} 0 & -1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix} \quad \Longrightarrow \quad ||C_J||_1 = ||C_J||_\infty = max(1/2, 1, 1/2) = 1\ .$$

(b) Por conseguinte,

$$\begin{aligned} e^{(0)} = x - x^{(0)} = (0.083, 0.033, -0.083) &\Longrightarrow \|e^{(0)}\|_1 = 0.199 \\ e^{(1)} = x - x^{(1)} = (-0.017, 0.083, 0.017) &\Longrightarrow \|e^{(1)}\|_1 = 0.117\ . \end{aligned}$$

Os resultados obtidos mostram que $x^{(1)}$ está mais próximo da solução exacta do que a aproximação inicial $x^{(0)}$. Acontece que $||C_J||_1 = 1$, pelo que para esta norma, ou para a norma $||\cdot||_\infty$, as majorações de erro (3.84) não são aplicáveis. No entanto, tal circunstância não permite concluir se o método converge ou não para a solução do sistema dado, uma vez que as referidas condições do Teorema do ponto fixo são apenas condições suficientes de convergência. Uma condição necessária e suficiente de convergência de métodos do tipo (3.83) será analisada adiante. ⧫

3.3.4 Método de Gauss-Seidel

O método de Gauss-Seidel[19] é um dos métodos iterativos mais comuns para resolução aproximada de sistemas lineares. Para deduzirmos a sua função iteradora, partimos de novo do sistema na forma (3.85), pág. 130.

As *fórmulas computacionais* deste método são as seguintes:

$$\begin{aligned}
x_1^{(k+1)} &= \frac{b_1 - a_{12}\, x_2^{(k)} - a_{13}\, x_3^{(k)} - \cdots - a_{1n}\, x_n^{(k)}}{a_{11}} \\
x_2^{(k+1)} &= \frac{b_2 - a_{21}\, x_1^{(k+1)} - a_{23}\, x_3^{(k)} - \cdots - a_{2n}\, x_n^{(k)}}{a_{22}} \qquad , \qquad k = 0, 1, 2, \ldots \\
&\vdots \\
x_n^{(k+1)} &= \frac{b_n - a_{n1}\, x_1^{(k+1)} - a_{n2}\, x_2^{(k+1)} - \cdots - a_{n,n-1}\, x_{n-1}^{(k+1)}}{a_{nn}}
\end{aligned} \tag{3.89}$$

Uma diferença em relação ao método de Jacobi consiste em que para se determinar a componente $x_i^{(k+1)}$ da iterada $(k+1)$ (com $i > 1$), utilizamos as componentes $x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}$ dessa mesma iterada, enquanto que no método de Jacobi as componentes de $x^{(k+1)}$ são calculadas apenas a partir das componentes de $x^{(k)}$ (da iterada anterior).

As expressões (3.89) podem ser escritas na forma

$$x_i^{(k+1)} = \frac{b_i - \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right)}{a_{ii}}, \quad i = 1 : n, \quad k = 0, 1, 2, \ldots \ . \tag{3.90}$$

Note que neste caso, ao contrário do método de Jacobi, a determinação da matriz de iteração C_{GS} e vector de correcção g_{GS} deixam de ser imediatos. No entanto, este método possui a vantagem computacional de permitir poupança de posições de memória do computador, visto que as iteradas consecutivas $x^{(1)}$, $x^{(2)}$, etc, podem ocupar as posições de memória do vector inicial $x^{(0)}$. Pelo contrário,

[19] Philipp Ludwig von Seidel, 1821-1896, matemático alemão.

no método de Jacobi, em cada iteração $k \geq 1$, é necessário manter posições de memória para os vectores $x^{(k-1)}$ e $x^{(k)}$. Além disso, em geral (embora nem sempre) o método de Gauss-Seidel, se convergente, converge mais rapidamente do que o método de Jacobi.

Exemplo 3.8. *Consideremos de novo o sistema (3.88), pág. 131.*

(a) Efectuar uma iteração do método de Gauss-Seidel, tomando como aproximação inicial $x^{(0)} = (0.5, 0.8, 1)$.

(b) Sabendo que a solução exacta do sistema é $x = (0.583, 0.833, 0.917)$, *calcular* $\|e^{(0)}\|_1$ *e* $\|e^{(1)}\|_1$.

(a) As fórmulas computacionais do método de Gauss-Seidel aplicado ao sistema, escrevem-se

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{2 - x_2^{(k)}}{2} \\
x_2^{(k+1)} &= \frac{2 + x_1^{(k+1)} - x_3^{(k)}}{2} = \frac{2 + \dfrac{2 - x_2^{(k)}}{2} - x_3^{(k)}}{2} = \frac{6 - x_2^{(k)} - 2\,x_3^{(k)}}{4}, \quad k = 0, 1, \ldots \\
x_3^{(k+1)} &= \frac{1 + x_2^{(k+1)}}{2} = \frac{1 + \dfrac{6 - x_2^{(k)} - 2\,x_3^{(k)}}{4}}{2} = \frac{10 - x_2^{(k)} - 2\,x_3^{(k)}}{8} \ .
\end{aligned}
$$

Assim, a respectiva matriz de iteração é

$$
C_{GS} = \begin{bmatrix} 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \\ 0 & -1/8 & -1/4 \end{bmatrix},
$$

e

$$
\begin{aligned}
||C_{GS}||_1 &= \max(0, 7/8, 3/4) = 7/8 < 1 \\
||C_{GS}||_\infty &= \max(1/2, 3/4, 3/8) = 3/4 < 1 \ .
\end{aligned}
$$

Atendendo ao teorema do ponto fixo, podemos garantir que o método converge para a solução $x = A^{-1}\,b$, qualquer que seja a escolha que fizermos da aproximação inicial $x^{(0)}$, em particular fazendo $x^{(0)} = (0.5, 0.8, 1)$. Por exemplo, na Figura 3.1 mostra-se a posição das primeiras 4 iteradas começando com $x^{(0)} = (0, 0, -7)$.

A primeira iterada do método $x^{(1)}$ tem as seguintes componentes:

$$
\begin{aligned}
x_1^{(1)} &= \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}}{a_{11}} = \frac{1}{2}(2 - 0.8 - 0) = 0.6 \\
x_2^{(1)} &= \frac{b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)}}{a_{22}} = \frac{1}{2}(2 + 0.6 - 1) = 0.8 \\
x_3^{(1)} &= \frac{b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)}}{a_{33}} = \frac{1}{2}(1 - 0 + 0.8) = 0.9 \ .
\end{aligned}
\tag{3.91}
$$

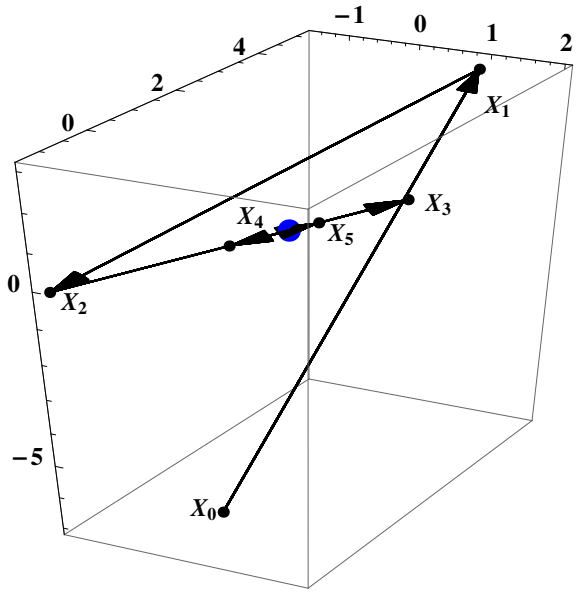


Figura 3.1: Ver Exemplo 3.8. Mostram-se 4 iteradas do método de Gauss-Seidel começando em $x^{(0)} = (0, 0, -7)$. O ponto de maiores dimensões representa a solução do sistema.

(b) Para os respectivos erros, obtemos

$$\begin{array}{ll} e^{(0)} = x - x^{(0)} = (0.083, 0.033, -0.083), & \|e^{(0)}\|_1 = 0.199 \\ e^{(1)} = x - x^{(1)} = (-0.017, 0.033, 0.017), & \|e^{(1)}\|_1 = 0.067\ . \end{array} \tag{3.92}$$

Tal como acontecia no caso do método de Jacobi, também aqui a norma do erro diminui da aproximação inicial para a primeira iterada, o que significa que esta está mais próxima da solução exacta do sistema, conforme se pode constatar observando a Figura 3.1. ♦

3.3.5 Forma matricial dos métodos iterativos

O estudo da convergência dos métodos iterativos para sistemas lineares é facilitado traduzindo esses métodos na forma matricial, tal como se descreve a seguir no caso dos métodos de Jacobi e de Gauss-Seidel.

Dada uma certa matriz A, começamos por definir as matrizes L, D, e U, tais que

$$L = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix},$$
$$\text{e} \qquad U = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \vdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}. \tag{3.93}$$

Obviamente, $A = L+D+U$. Supomos que todas as entradas diagonais da matriz A são diferentes de zero, ou seja,

$$a_{ii} \neq 0, \qquad i = 1 : n.$$

Assumimos, portanto, que a matriz D é invertível. Por isso se diz que a soma $A = D + (L + U)$ corresponde a uma *decomposição regular* da matriz A, no sentido em que a primeira parcela da soma referida, D, é uma matriz (facilmente) invertível.

Método de Jacobi na forma matricial

Utilizando as matrizes L, D e U introduzidas em (3.93), vejamos como se pode escrever a fórmula iterativa (3.83), pág. 128, do método de Jacobi, identificando o vector g_J e a matriz de iteração C_J correspondentes.

Comecemos por escrever a fórmula (3.87) recorrendo às matrizes L, D e U,

$$x^{(k+1)} = D^{-1}\left(b - L\,x^{(k)} - Ux^{(k)}\right),$$

ou, equivalentemente,

$$x^{(k+1)} = D^{-1}b - D^{-1}(L + U)\,x^{(k)}.$$

Comparando esta última igualdade com a fórmula geral para os métodos iterativos (3.83), pág. 128, concluimos que no caso do método de Jacobi o vector auxiliar g_J e a matriz de iteração têm a forma,

$$C_J = -D^{-1}(L + U), \qquad g_J = D^{-1}b\ . \tag{3.94}$$

Uma vez que todas as entradas da diagonal da matriz D são não nulas[20], a matriz

[20]Se a diagonal principal da matriz do sistema dado possuir alguma entrada nula, deverá começar-se por reordenar as equações de modo que o sistema resultante possua todas as entradas da diagonal principal não nulas.

inversa D^{-1} pode ser determinada imediatamente,

$$D^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{a_{nn}} \end{bmatrix}.$$

Por conseguinte, a matriz de iteração tem a forma (que já conhecíamos),

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{bmatrix}. \tag{3.95}$$

Relembre-se que no caso do método de Jacobi, tanto a matriz de iteração C_J, como o vector de correcção g_J, podem ser obtidos imediatamente a partir das fórmulas computacionais (3.86), pág. 130.

Método de Gauss-Seidel na forma matricial

Vejamos agora como se pode traduzir o processo do método de Gauss-Seidel na forma (3.83), pág. 128.

Com o auxílio das matrizes L, D e U, a fórmula (3.90) pode escrever-se como

$$x^{(k+1)} = D^{-1}\left(b - L\,x^{(k+1)} - U\,x^{(k)}\right). \tag{3.96}$$

Multiplicando por D ambos os membros de (3.96), obtém-se

$$D\,x^{(k+1)} = b - L\,x^{(k+1)} - U\,x^{(k)}. \tag{3.97}$$

Passando para o primeiro membro os termos que contêm $x^{(k+1)}$, resulta

$$(L+D)\,x^{(k+1)} = b - U\,x^{(k)}.$$

Uma vez que a matriz D é invertível, $L+D$ também o é (o determinante de $L+D$ é igual ao determinante de D). Assim, podemos escrever

$$x^{(k+1)} = (L+D)^{-1}b - (L+D)^{-1}Ux^{(k)}. \tag{3.98}$$

Finalmente, comparando a equação (3.98) com a fórmula geral para os métodos iterativos, concluímos que a respectiva matriz de iteração e o vector auxiliar têm a forma

$$C_{GS} = -(L+D)^{-1}U, \qquad g_{GS} = (L+D)^{-1}b \tag{3.99}$$

Em geral não é possível encontrar uma forma explícita para a inversa de $(L+D)$. Tudo o que se pode dizer é tratar-se de uma matriz triangular inferior onde os seus elementos diagonais são os inversos dos elementos diagonais de A. Logo, também não é possível encontrar uma forma imediatamente explícita para a matriz de iteração C_{GS}.

Podemos no entanto concluir que a matriz C_{GS} possui a primeira coluna com entradas nulas (no método de Jacobi a respectiva matriz de iteração possui diagonal principal de entradas nulas).

Exemplo 3.9. *Determinemos respectivamente o vector de correcção e a matriz de iteração dos métodos de Jacobi e de de Gauss-Seidel, para o sistema do Exemplo 3.3.3, pág. 131.*

Para o método de Jacobi,

$$g_J = D^{-1}b = \left(\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \frac{b_3}{a_{33}}\right)^T = \left(1, 1, \frac{1}{2}\right)^T.$$

A matriz C_J obtém-se a partir de (3.95),

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -1/2 & 0 \\ 1/2 & 0 & -1/2 \\ 0 & 1/2 & 0 \end{bmatrix}.$$

Podemos no entanto obter C_J e g_J directamente a partir das fórmulas computacionais para este método que resultam imediatamente da rescrita do sistema dado na forma de ponto fixo $x = C\,x + d$.

No caso do método de Gauss-Seidel, para poder determinar o vector g_{GS} e a matriz de iteração começamos por calcular a matriz inversa de $L + D$:

$$(L+D)^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/4 & 1/2 & 0 \\ 1/8 & 1/4 & 1/2 \end{bmatrix}.$$

Das fórmulas (3.99) obtém-se,

$$C_{GS} = \begin{bmatrix} 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \\ 0 & -1/8 & -1/4 \end{bmatrix}, \qquad g_{GS} = \left(1, \frac{3}{2}, \frac{5}{4}\right).$$

♦

3.3.6 Convergência

Uma vez definido um método iterativo para calcular aproximações da solução de um sistema linear, é fundamental saber em que condições esse método gera uma sucessão que converge para essa solução. Nos teoremas adiante estabelecem-se condições sobre a matriz do sistema que garantem a convergência dos métodos iterativos considerados.

Resulta das fórmulas (3.82) e (3.83), pág. 128, que os erros das iteradas satisfazem as seguintes igualdades fundamentais,

$$e^{(k+1)} = x - x^{(k+1)} = C\,(x - x^{(k)}), \qquad k = 0, 1, 2, \ldots \tag{3.100}$$

isto é,

$$e^{(k+1)} = C\,e^{(k)}, \qquad k = 0, 1, 2, \ldots \tag{3.101}$$

onde C é a matriz de iteração do método considerado.

No parágrafo anterior já foi analisada a forma das matrizes de iteração dos métodos de Jacobi e Gauss-Seidel. Vejamos agora quais as propriedades da matriz C que garantem convergência de um método iterativo desta natureza.

Em primeiro lugar, notemos que da igualdade (3.101) resulta imediatamente uma relação que exprime o erro de qualquer iterada através do erro da aproximação inicial:

$$e^{(k)} = C^k\,e^{(0)}, \qquad k = 0, 1, 2, \ldots\ . \tag{3.102}$$

A relação (3.102) suscita naturalmente a seguinte definição.

Definição 3.10. Uma matriz $C \in \mathbb{R}^{n\times n}$, diz-se *convergente* se e só se

$$\lim_{k\to\infty} C^k\,x = 0, \qquad \forall x \in \mathbb{R}^n. \tag{3.103}$$

Estamos agora em condições de enunciar um teorema que fornece uma condição necessária e suficiente para a convergência dos métodos iterativos do tipo (3.83), pág. 128.

Teorema 3.5. Seja $(x_k)_{k\geq 0}$ uma sucessão em $\mathbb{R}^n$, gerada pela fórmula (3.83), onde C é uma matriz de iteração associada ao sistema $A\,x = b$. A sucessão $(x_k)_{k\geq 0}$ converge para a solução do sistema, qualquer que seja a aproximação inicial $x^{(0)}$, se e só se a matriz C for convergente.

Demonstração. (Condição suficiente).
Seja C uma matriz convergente, e $e^{(k)}$ o erro da k-ésima iterada. De acordo com as fórmulas (3.102) e (3.103), temos

$$\lim_{k\to\infty} e^{(k)} = \lim_{k\to\infty} C^k\,e^{(0)} = 0, \tag{3.104}$$

qualquer que seja o vector $e^{(0)} \in \mathbb{R}^n$, independentemente da norma considerada. Isto significa que o método iterativo converge, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbb{R}^n$.

(Condição necessária). Suponhamos que a matriz C não é convergente. Então, existe um vector $v \in \mathbb{R}^n$, tal que a sucessão $(C^k\, v)_{k\geq 0}$ não converge para o vector nulo. Seja $x^{(0)} = x + v$, onde x é a solução exacta do sistema. De acordo com (3.102), temos $e^{(k)} = C^k\, v$ e, por definição de v, a sucessão $(e^{(k)})_{k\geq 0}$ não tende para o vector nulo, significando que o método iterativo não é convergente, se tomarmos como aproximação inicial $x^{(0)} = x + v$. □

Em geral não é fácil averiguar se a matriz C é ou não convergente usando directamente a Definição 3.9. Vamos a seguir apresentar dois teoremas que nos permitem decidir sobre a convergência de uma matriz.

Teorema 3.6. Seja $C \in \mathbb{R}^{n\times n}$. Se numa dada norma matricial $||\cdot||_M$, induzida por uma norma vectorial $||\cdot||_V$, se verificar

$$\|C\|_M < 1,$$

então a matriz C é convergente.

Demonstração. Seja x um vector arbitrário de $\mathbb{R}^n$. De acordo com a propriedade submultiplicativa das normas matriciais, referida no parágrafo 3.0.1, pág. 93, temos

$$\|C^k\, x\|_V \leq \|C^k\|_M\, \|x\|_V \leq (\|C\|_M)^k\, \|x\|_V\ . \tag{3.105}$$

Das desigualdades (3.105) resulta imediatamente que, sendo $\|C\|_M < 1$,

$$\lim_{k\to\infty} \|C^k\, x\|_V = 0,$$

o que significa, por definição, que a matriz C é convergente. □

Fixada uma norma vectorial e a correspondente norma matricial induzida, uma vez que o erro de uma iterada k de um determinado método iterativo convergente, de matriz C, satisfaz a condição (3.102), quanto menor for a norma $||C||$, com $||C|| < 1$, mais depressa o método convergirá para a solução do sistema linear em causa. Entre dois métodos distintos aplicados a um sistema $A\, x = b$, cujas normas da respectiva matriz de iteração tenham valores diferentes e inferiores a um, o método de convergência mais rápida (para essa norma) será aquele cuja matriz de iteração tenha o valor da norma menor.

Pode acontecer que para uma determinada norma se tenha $||C|| \geq 1$ e no entanto a matriz de iteração ser convergente. O resultado fundamental a seguir dá-nos uma condição necessária e suficiente de convergência da matriz de iteração.

Teorema 3.7. Para que a matriz $C \in \mathbb{R}^{n\times n}$ seja convergente é necessário e suficiente que o seu raio espectral $\rho(C)$ satisfaça a condição

$$\rho(C) < 1 \ . \tag{3.106}$$

Demonstração. (Condição suficiente). Se tivermos $\rho(C) = \rho < 1$, de acordo com [18], p. 12, para qualquer $\epsilon > 0$, existe uma norma matricial $N(\epsilon)$ tal que

$$\|C\|_{N(\epsilon)} \leq \rho + \epsilon \ .$$

Se considerarmos $\epsilon = \dfrac{1-\rho}{2}$, obtemos

$$\|C\|_{N(\epsilon)} \leq \frac{\rho+1}{2} < 1 \ . \tag{3.107}$$

Da desigualdade (3.107) resulta, pelo Teorema 3.5, que a matriz C é convergente.

(Condição necessária). Suponhamos que a condição (3.106) não se verifica, isto é, que $\rho(C) \geq 1$. Então, existe pelo menos um valor próprio λ de C, tal que $|\lambda| = \rho \geq 1$. Seja v um vector próprio de C, associado ao valor próprio λ. Logo, para qualquer norma vectorial, verifica-se

$$\|C^k\, v\| = \|\lambda^k\, v\| = |\lambda|^k\, \|v\| \ . \tag{3.108}$$

Visto que $|\lambda| = \rho \geq 1$, resulta de (3.108) que a sucessão $(C^k\, v)_{k\geq 0}$ não converge para o vector nulo, pelo que a matriz C não é convergente. □

Se dispusermos de informação a respeito do raio espectral das matrizes de iteração de dois métodos iterativos distintos, aplicados a uma sistema $A\,x = b$, o método de convergência mais rápida será aquele cuja matriz de iteração tenha um raio espectral menor.

3.3.7 Critérios de convergência

Com base nos Teoremas 3.5 e 3.6, podemos obter critérios de convergência para os métodos de Jacobi e de Gauss-Seidel sem necessitarmos de recorrer ao raio espectral da respectiva matriz de iteração. Comecemos por introduzir algumas definições.

Definição 3.11. (Dominância estrita por linhas ou colunas)
Diz-se que a matriz $A \in \mathbb{R}^{n\times n}$ é de *diagonal estritamente dominante por linhas*, se forem satisfeitas as condições

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| \; < \; |a_{ii}|, \qquad i = 1:n. \tag{3.109}$$

A matriz A diz-se de *diagonal estritamente dominante por colunas*, se

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| \; < \; |a_{jj}|, \qquad j = 1:n. \tag{3.110}$$

Embora com abuso de linguagem, diremos de modo abreviado que uma matriz $A \in \mathbb{R}^{n\times n}$ é *estritamente dominante* se for de diagonal estritamente dominante por linhas, ou por colunas. Uma matriz estritamente dominante é necessariamente não singular.

Proposição 3.1. Se a matriz $A \in \mathbb{R}^{n\times n}$ é de diagonal estritamente dominante, então A é não singular.

Demonstração. Suponhamos que a matriz A é de diagonal estritamente dominante por linhas e singular. Assim, $\lambda = 0$ é valor próprio de A. Seja $v \neq 0$ vector próprio pertencente a $\lambda = 0$, isto é,

$$A\,v = \lambda\,v = 0\ .$$

A linha i da igualdade $A\,v = 0$, escreve-se

$$\sum_{j=1}^{n} a_{ij}\,v_j = 0 \Longleftrightarrow a_{ii}\,v_i = -\sum_{j=1, j\neq i}^{n} a_{ij}\,v_j, \qquad i = 1:n. \tag{3.111}$$

Seja l o primeiro índice para o qual

$$|v_l| = \max_{1\leq i\leq n} |v_i|, \quad \text{com} \quad |v_l| \neq 0 \quad \text{pois} \quad v \neq 0. \tag{3.112}$$

Fazendo $i = l$ em (3.111), tem-se

$$a_{ll}\,v_l = -\sum_{j=1, j\neq l}^{n} a_{lj}\,v_j,$$

donde

$$|a_{ll}|\,|v_l| \leq \sum_{j=1, j\neq l}^{n} |a_{lj}|\,|v_j| \; \leq |v_l| \sum_{j=1, j\neq l}^{n} |a_{lj}| \quad .$$

A última desigualdade é válida atendendo a (3.112). Logo,

$$|a_{ll}| \leq \sum_{j=1, j\neq l}^{n} |a_{lj}|,$$

desigualdade falsa, porquanto por hipótese a matriz A é de diagonal estritamente dominante por linhas. Conclui-se, portanto, que A é não singular.

No caso da matriz A ser de diagonal estritamente dominante por colunas, resulta que a matriz transposta A^T é de diagonal estritamente dominante por linhas. O resultado anterior garante que A^T é não singular e, consequentemente, A é também não singular. □

Os métodos de Jacobi e de Gauss-Seidel são convergentes quando aplicados a um sistema cuja matriz dos coeficientes seja de diagonal estritamente dominante, conforme se mostra no Teorema 3.8 adiante. Comecemos por demonstrar o seguinte resultado preliminar.

Proposição 3.2. Seja $A \in \mathbb{R}^{n\times n}$ matriz de diagonal estritamente dominante (por linhas ou por colunas) e $A = D + L + U$ uma sua decomposição regular. Conside-se $\mu \in \mathbb{C}$ e

$$\begin{array}{l} A_\mu = \mu\, D + L + U, \quad \text{onde} \quad |\mu| \geq 1 \\ A'_\mu = \mu\,(D + L) + U, \quad \text{onde} \quad |\mu| \geq 1\ . \end{array} \tag{3.113}$$

As matrizes A_μ e A'_μ são de diagonal estritamente dominante (por linhas ou por colunas).

Corolário 3.1. Nas condições da Proposição 3.2, as matrizes A_μ e A'_μ são não singulares.

Demonstração. Suponhamos que a matriz A é estritamente diagonal dominante por linhas (o caso da dominância estrita por colunas pode mostrar-se de modo análogo e é deixado como exercício).

As entradas da diagonal principal das matrizes A_μ e A'_μ são $\mu\, a_{ii}$, para $i = 1 : n$. Atendendo à hipótese de dominância estrita da matriz A, tem-se

$$|\mu\, a_{ii}| = |\mu|\, |a_{ii}| > |\mu| \sum_{j=1, \jmath\neq i}^{n} |a_{ij}|,$$

isto é,

$$|\mu\, a_{ii}| > |\mu| \left(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \right).$$

Ou seja,

$$|\mu\, a_{ii}| > |\mu| \sum_{j=1}^{i-1} |a_{ij}| + |\mu| \sum_{j=i+1}^{n} |a_{ij}| \ . \tag{3.114}$$

A desigualdade (3.114) permite-nos concluir dominância estrita, por linhas, das matrizes A_μ e A'_μ. Com efeito, por hipótese tem-se $|\mu| \geq 1$, logo
(i)

$$|\mu\, a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|, \qquad i = 1 : n,$$

o que significa que A_μ possui diagonal estritamente dominante por linhas.
(ii)

$$|\mu\, a_{ii}| > |\mu| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|, \qquad i = 1 : n,$$

ou seja, a matriz A'_μ é também de diagonal estritamente dominante por linhas.

O Corolário 3.1 resulta imediatamente da Proposição 3.1.

□

Associadas às matrizes D, L e U definindo a decomposição regular de A, é útil considerar as matrizes triangulares (com zeros na respectiva diagonal principal)

$$\begin{aligned} L_1 &= D^{-1}\, L \\ U_1 &= D^{-1}\, U \ . \end{aligned} \tag{3.115}$$

Podemos agora enunciar o resultado que nos garante convergência dos métodos de Jacobi e de Gauss-Seidel, quando aplicados a sistemas de matriz dos coeficientes estritamente dominante.

Teorema 3.8. Dado o sistema $A\, x = b$, onde $A \in \mathbb{R}^{n\times n}$ é matriz de diagonal estritamente dominante (por linhas ou por colunas), os métodos de Jacobi e de Gauss-Seidel são convergentes para a solução $x = A^{-1}\, b$, qualquer que seja a aproximação inicial $x^{(0)}$ escolhida.

Se para alguma norma matricial induzida se verifica a desigualdade

$$||L_1|| + ||U_1|| < 1, \tag{3.116}$$

então

$$||C_J|| = ||L_1 + U_1|| < 1 \tag{3.117}$$

e

$$||C_{GS}|| \leq \frac{||U_1||}{1 - ||L_1||} < 1, \tag{3.118}$$

onde L_1 e U_1 são as matrizes triangulares (3.115).

Demonstração. Método de Jacobi

A matriz de iteração é $C_J = -D^{-1}(L+U)$. Seja $\lambda \in Sp(C_J)$. A equação característica $\det(\lambda I - C_J) = 0$, pode escrever-se como

$$\begin{aligned}\det(\lambda I + D^{-1}(L+U)) &= \det(\lambda D^{-1} D + D^{-1}(L+U)) \\ &= \det(D^{-1}(\lambda D + L + U)) \\ &= \det(D^{-1}) \times \det(A_\lambda) = 0 \ .\end{aligned}$$

Visto que a matriz D é não singular, a última igualdade implica que $\det(A_\lambda) = 0$, isto é, que A_λ seja singular. Atendendo à Proposição 3.2, pág. 142, para $\mu = \lambda$, a singularidade de A_λ só é possível caso $|\lambda| < 1$. Por conseguinte, $\rho(C_J) < 1$, o que implica convergência do método para a solução do sistema.

Mostremos que sob a condição (3.116) é satisfeita a desigualdade (3.117). Fixada uma norma vectorial em $\mathbb{R}^n$, seja $x \in \mathbb{R}^n$ tal que $||x|| = 1$. Fazendo

$$y = C_J\, x = -D^{-1}(L+U)\, x = -(L_1 + U_1)\, x,$$

resulta, por aplicação da desigualdade triangular,

$$||y|| \leq ||L_1 + U_1|| \leq ||L_1|| + ||U_1|| \ .$$

Por conseguinte,

$$||C_J|| = max_{||x||=1} ||C_J\, x|| \leq ||L_1|| + ||U_1|| < 1 \ .$$

Método de Gauss-Seidel

Uma vez que a matriz de iteração do método é $C_{GS} = -(D+L)^{-1} U$, a respectiva equação característica $\det(\lambda I - C_{GS}) = 0$, pode escrever-se como

$$\begin{aligned}\det(\lambda I + (D+L)^{-1} U) &= \det(\lambda (D+L)^{-1}(D+L) + (D+L)^{-1} U) \\ &= \det((D+L)^{-1}(\lambda (D+L) + U)) \, .\end{aligned}$$

Assim,

$$\det\left((D+L)^{-1}\right) \times \det(A'_\lambda) = 0 \ .$$

Como a matriz $D+L$ é não singular, a igualdade anterior implica que $\det(A'_\lambda) = 0$, isto é, a matriz A'_λ é singular. Fazendo $\mu = \lambda$ na Proposição 3.2, conclui-se que necessariamente

$$|\lambda| < 1 \Longrightarrow \rho(C_{GS}) < 1,$$

logo o método converge.

Mostremos a validade da desigualdade (3.118). Fixada uma norma vectorial, seja $x \in \mathbb{R}^n$ tal que $||x|| = 1$. De

$$y = C_{GS}\, x = -(D+L)^{-1} U\, x,$$

obtém-se

$$(D+L)\,y = -U\,x \Longleftrightarrow D\,y = -L\,y - U\,x \Longleftrightarrow y = \text{Ł}_1\,y - U_1\,x\ .$$

Assim,

$$||y|| \leq ||L_1||\,||y|| + ||U_1|| \Longleftrightarrow (1 - ||L_1||)\,||y|| \leq ||U_1||\ .$$

Sob a hipótese (3.116), tem-se que $||L_1|| < 1$ e

$$||C_{GS}|| = max_{||x||=1}||y|| \leq \frac{||U_1||}{1 - ||L_1||} < 1\ .$$

□

Exemplo 3.10. *Pretende-se aplicar os métodos de Jacobi e de Gauss-Seidel a dois sistemas lineares cuja matriz dos coeficientes é, respectivamente,*

$$(i)\quad A = \begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix} \quad (ii)\quad A = \begin{bmatrix} 3 & -2 \\ 1 & 3 \end{bmatrix}.$$

Uma vez que ambas as matrizes são estritamente dominantes (por linhas e/ou por colunas), o Teorema 3.8 garante que ambos os métodos são convergentes para a solução de cada um dos sistemas considerados, independentemente da aproximação inicial $x^{(0)}$ escolhida.

Fixada a norma $||\cdot||_\infty$, é verdade que o método de Gauss-Seidel converge mais rapidamente do que o método de Jacobi?

Comecemos por mostrar que as relações (3.116)–(3.118), pág. 143, são aplicáveis ao sistema de matriz (i) mas não se aplicam ao sistema de matriz (ii). Além disso, iremos verificar que

$$||C_J||_\infty = ||C_{GS}||_\infty = 2/3.$$

Conclui-se destas igualdades que ambos os métodos convergem. No entanto a informação quanto à norma da matriz de iteração de cada um dos métodos, por terem o mesmo valor, não nos permite decidir qual dos dois métodos irá convergir mais rapidamente. Para esse efeito teremos necessidade de comparar o raio espectral $\rho(C_J)$ com o raio espectral $\rho(C_{GS})$.

Matriz (i)

$$L_1 = D^{-1}\,L = \begin{bmatrix} 0 & 0 \\ -1/3 & 0 \end{bmatrix}, \qquad U_1 = D^{-1}\,U = \begin{bmatrix} 0 & 1/3 \\ 0 & 0 \end{bmatrix}$$

$$C_J = -(L_1 + U_1) = \begin{bmatrix} 0 & -1/3 \\ 1/3 & 0 \end{bmatrix}.$$

Assim,

$$||L_1||_\infty = 1/3, \quad ||U_1||_\infty = 1/3, \quad ||C_J||_\infty = 1/3 \ .$$

É verdade que

$$||C_J||_\infty \leq ||L_1||_\infty + ||U_1||_\infty = 2/3 < 1.$$

Tem-se

$$\det(\lambda\, I - C_J) = 0 \Longleftrightarrow \lambda^2 + 1/9 = 0 \Longrightarrow \rho(C_J) = 1/3 \ .$$

Passemos ao método de Gauss-Seidel:

$$\begin{aligned} C_{GS} = -(D+L)^{-1}\, U \ &= -\begin{bmatrix} 3 & 0 \\ -1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\ &= -\frac{1}{9}\begin{bmatrix} 3 & 0 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1/3 \\ 0 & -1/9 \end{bmatrix} \\ &\Longrightarrow ||C_{GS}||_\infty = 1/3 = ||C_J||_\infty \ . \end{aligned}$$

Note-se que são válidas as desigualdades

$$||C_{GS}||_\infty \leq \frac{||U_1||_\infty}{1 - ||L_1||_\infty} = \frac{1}{2} < 1 \ .$$

Dado que C_{GS} é triangular superior, o seu raio espectral obtém-se imediatamente, sendo $\rho(C_{GS}) = 1/9$. Uma vez que este valor é inferior ao valor de $\rho(C_J)$, conclui-se que o método de Gauss-Seidel converge mais rapidamente do que o método de Jacobi.

Matriz (ii)

$$L_1 = D^{-1}\, L = \begin{bmatrix} 0 & 0 \\ 1/3 & 0 \end{bmatrix}, \qquad U_1 = D^{-1}\, U = \begin{bmatrix} 0 & -2/3 \\ 0 & 0 \end{bmatrix}$$

$$C_J = -(L_1 + U_1) = \begin{bmatrix} 0 & -2/3 \\ 1/3 & 0 \end{bmatrix}.$$

Assim,

$$||L_1||_\infty = 1/3, \quad ||U_1||_\infty = 2/3, \quad ||C_J||_\infty = 2/3 < 1 \ .$$

Neste caso

$$||L_1||_\infty + ||U_1||_\infty = 1,$$

pelo que a condição (3.116), pág. 143, não é aplicável. Como $\det(\lambda\, I - C_J) = \lambda^2 + 2/9 = 0 \Longrightarrow \rho(C_J) = \sqrt{2}/3 < 1$, logo o método converge.

Para o método de Gauss-Seidel, tem-se

$$\begin{aligned} C_{GS} = -(D+L)^{-1}U &= -\begin{bmatrix} 3 & 0 \\ 1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -2 \\ 0 & 0 \end{bmatrix} \\ &= -\frac{1}{9}\begin{bmatrix} 3 & 0 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 0 & -2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2/3 \\ 0 & -2/9 \end{bmatrix} \\ &\Longrightarrow ||C_{GS}||_\infty = 2/3 = ||C_J||_\infty \ . \end{aligned}$$

Dado que C_{GS} é triangular superior, o seu raio espectral obtém-se imediatamente, sendo $\rho(C_{GS}) = 2/9 < \rho(C_J)$. Por conseguinte conclui-se que o método de Gauss-Seidel converge mais rapidamente do que o método de Jacobi.

Convergência do método de Jacobi

Vamos particularizar o resultado obtido no Teorema 3.8 escolhendo normas matriciais induzidas apropriadas, quando a matriz A é estritamente dominante. Por exemplo, para a norma $||\cdot||_\infty$, resulta o seguinte critério de convergência para o método de Jacobi.

Teorema 3.9. Se a matriz A for de diagonal estritamente dominante por linhas, então o método de Jacobi converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbb{R}^n$.

Demonstração. Sendo a matriz A de diagonal estritamente dominante por linhas, das desigualdades (3.109), resulta

$$\sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1, \qquad i = 1:n. \tag{3.119}$$

De acordo com a forma da matriz C_J, dada por (3.94), pág. 135, as desigualdades (3.119) implicam

$$\|C_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1. \tag{3.120}$$

Atendendo ao Teorema 3.6, a condição (3.120) garante que a matriz C_J é convergente. De acordo com o Teorema 3.5, o método de Jacobi é convergente, qualquer que seja a aproximação inicial. □

No caso de a matriz A ser de diagonal estritamente dominante por colunas, pode considerar-se a norma induzida definida a seguir.

Teorema 3.10. Se a matriz A é de diagonal estritamente dominante por colunas, então o método de Jacobi converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbb{R}^n$.

Demonstração. Suponhamos que a matriz A satisfaz (3.110) e que D é a matriz diagonal (invertível) cujas entradas da diagonal principal são as de A, isto é, $D = diag(a_{11}, \ldots, a_{nn})$. Podemos definir uma norma matricial $||\cdot||_M$,

$$\|X\|_M \;=\; \|D\,X\,D^{-1}\|_1, \qquad \forall\, X \in \mathbb{R}^{(n\times n)}. \tag{3.121}$$

Das condições (3.110) obtém-se,

$$\|C_J\|_M \;=\; \|D\,C_J\,D^{-1}\|_1 \;= ||-(L+U)\,D^{-1}||_1 < 1 \quad . \tag{3.122}$$

De acordo com o Teoremas 3.5 e 3.6, da desigualdade (3.122) resulta que o método de Jacobi converge para a solução do sistema $A\,x \;=\; b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbb{R}^n$. □

Exemplo 3.11. *(a) A matriz A do sistema do Exemplo 3.7, pág. 131, é da forma*

$$A \;=\; \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}. \tag{3.123}$$

Se aplicarmos o método de Jacobi a um sistema $A\,x = b$, com b qualquer, poderemos garantir convergência desse método?

(b) Pode-se garantir que o método de Jacobi converge quando A é a matriz a seguir?

$$A \;=\; \begin{bmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 0 & 0 & 2 \end{bmatrix}. \tag{3.124}$$

(a) Verifica-se facilmente que a matriz não é de diagonal estritamente dominante por linhas, uma vez que,

$$|a_{22}| = |a_{21}| + |a_{23}| \;.$$

Do mesmo modo se pode verificar que A não tem a diagonal estritamente dominante por colunas. Por conseguinte, os Teoremas 3.9 e 3.10 não são aqui aplicáveis. Vejamos se é possível aplicar directamente o Teorema 3.7, pág. 140.

A matriz C_J tem a forma,

$$C_J \;=\; \begin{bmatrix} 0 & -1/2 & 0 \\ 1/2 & 0 & -1/2 \\ 0 & 1/2 & 0 \end{bmatrix}. \tag{3.125}$$

Os valores próprios de C_J são raízes da equação

$$\lambda^3 + \frac{\lambda}{2} \;=\; 0,$$

ou seja,

$$\lambda_1 = 0, \quad \lambda_2 = \frac{i}{\sqrt{2}}, \quad \lambda_3 = -\frac{i}{\sqrt{2}} .$$

Por conseguinte, o raio espectral de C_J é

$$\rho(C_J) = |\lambda_2| = \frac{1}{\sqrt{2}} < 1 .$$

Logo, pelo Teorema 3.7, podemos concluir que o método de Jacobi converge para a solução do sistema considerado, qualquer que seja a aproximação inicial.

(b) Para a matriz A em (3.124), a matriz de iteração C_J associada ao sistema $A\,x = b$, tem a forma

$$C_J = \begin{bmatrix} 0 & -1 & 0 \\ -1/3 & 0 & -1/3 \\ 0 & 0 & 0 \end{bmatrix} .$$

Tomando $D = diag(2, 3, 2)$, obtém-se

$$D\,C_J D^{-1} = -(L+U)\,D^{-1} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & -1/3 & 0 \\ -1/6 & 0 & -1/6 \\ 0 & 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -2/3 & 0 \\ -1/2 & 0 & -1/2 \\ 0 & 0 & 0 \end{bmatrix} .$$

A respectiva norma $||C_J||_M$, definida em (3.121), é

$$||C_J||_M = ||D\,C_J\,D^{-1}||_1 = \max(1/2, 2/3, 1/2) = 2/3 < 1,$$

pelo que podemos garantir convergência do método de Jacobi. Note que poderíamos chegar à mesma conclusão aplicando o Teorema 3.10. ♦

Convergência do método de Gauss-Seidel

Embora o Teorema 3.8, pág. 143, seja válido para os métodos de Jacobi e de Gauss-Seidel quando se verifica dominância estrita da matriz A (por linhas, por colunas, ou por linhas e colunas), vamos particularizar neste parágrafo apenas para dominância por linhas e para o método de Gauss-Seidel. Será fixada notação que nos permite estabelecer certas majorações de erro, notação essa que voltará a ser usada no parágrafo seguinte onde se comparam os métodos de Jacobi e de Gauss-Seidel quanto à rapidez de convergência.

Representemos por C_{GS} a matriz

$$C_{GS} = -(L+D)^{-1}U \ . \tag{3.126}$$

Segundo o Teorema 3.5, pág. 138, o método de Gauss-Seidel converge, qualquer que seja a aproximação inicial, se e só se a matriz C_{GS} for convergente. Para que tal ocorra, de acordo com o Teorema 3.7 é necessário e suficiente que o seu raio espectral seja menor do que 1.

Vamos mostrar que o método de Gauss-Seidel converge sempre que a matriz do sistema tiver a diagonal estritamente dominante por linhas.

Considerem-se, para $i = 1 : n$,

$$\alpha_i = \begin{cases} 0, & \text{se} \quad i = 1 \\ \sum_{j=1}^{i-1} \left| \dfrac{a_{ij}}{a_{ii}} \right|, & \text{se} \quad i = 2 : n; \end{cases} \qquad \beta_i = \begin{cases} 0, & \text{se} \quad i = n \\ \sum_{j=i+1}^{n} \left| \dfrac{a_{ij}}{a_{ii}} \right|, & \text{se} \quad i = 1 : (n-1) \; . \end{cases} \tag{3.127}$$

Sendo conhecidos α_i e β_i, defina-se a grandeza η através da fórmula

$$\eta = \max_{i=1,\ldots,n} \left(\frac{\beta_i}{1 - \alpha_i} \right). \tag{3.128}$$

Teorema 3.11. Seja A matriz de um sistema linear com *diagonal estritamente dominante por linhas*. O método de Gauss-Seidel converge, qualquer que seja a aproximação inicial, e é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \eta^k \, \|e^{(0)}\|_\infty. \tag{3.129}$$

Demonstração. Da fórmula (3.90), pág. 132, deduz-se facilmente que o erro da k-ésima iterada do método de Gauss-Seidel satisfaz a igualdade

$$e_i^{(k+1)} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} \, e_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} \, e_j^{(k)} \right), \qquad i = 1 : n, \quad k = 0, 1, \ldots \; . \tag{3.130}$$

Tomando o módulo de ambos os membros de (3.130), e entrando em conta com as definições das grandezas α_i e β_i, obtém-se

$$|e_i^{(k+1)}| \leq \alpha_i \, \|e^{(k+1)}\|_\infty + \beta_i \|e^{(k)}\|_\infty, \quad i = 1 : n, \quad k = 0, 1, \ldots \; . \tag{3.131}$$

Seja m o índice para o qual se verifica $|e_m^{(k+1)}| = \|e^{(k+1)}\|_\infty$. Então, escrevendo a desigualdade (3.131), com $i = m$, obtém-se

$$\|e^{(k+1)}\|_\infty \leq \alpha_m \, \|e^{(k+1)}\|_\infty + \beta_m \|e^{(k)}\|_\infty, \qquad k = 0, 1, \ldots \; .$$

ou, equivalentemente,

$$\|e^{(k+1)}\|_\infty (1 - \alpha_m) \leq \beta_m \|e^{(k)}\|_\infty, \qquad k = 0, 1, \ldots . \tag{3.132}$$

Visto que $\alpha_m < 1$, podemos dividir ambos os membros de (3.132) por $1 - \alpha_m$, e obter

$$\|e^{(k+1)}\|_\infty \leq \frac{\beta_m}{1 - \alpha_m} \|e^{(k)}\|_\infty \leq \eta \, \|e^{(k)}\|_\infty, \qquad k = 0, 1, \ldots \; . \tag{3.133}$$

Das desigualdades (3.133) resulta a estimativa de erro (3.129).

Por outro lado, uma vez que a matriz tem a diagonal estritamente dominante por linhas, $\eta < 1$. Logo, a desigualdade (3.129) implica que

$$\lim_{k\to\infty} \|e^{(k)}\|_\infty = 0,$$

o que garante a convergência do método de Gauss-Seidel, qualquer que seja a aproximação inicial. □

Exemplo 3.12. *Consideremos o mesmo sistema linear dos exemplos anteriores, com matriz*

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}.$$

Prove-se que o método de Gauss-Seidel converge quando aplicado a um sistema linear $A\,x = b$, com segundo membro b arbitrário.

A matriz A não é de diagonal estritamente dominante por linhas nem por colunas. Por conseguinte, o Teorema 3.11, pág. 150, não é aqui aplicável.

Vejamos se é possível aplicar directamente o Teorema 3.7, pág. 140. A matriz C_{GS}, de acordo com (3.3.5), tem a forma

$$C_{GS} = \begin{bmatrix} 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \\ 0 & -1/8 & -1/4 \end{bmatrix}. \tag{3.134}$$

Ora, como

$$||C_{GS}||_\infty = \max(1/2, 3/4, 3/8) = 3/4 < 1,$$

podemos garantir convergência do método. Uma vez que, para qualquer norma induzida, $\rho(C_{GS}) \leq ||C_{GS}||$ (ver Teorema 3.1, pág. 97), conclui-se que $\rho(C_{GS}) < 1$.

Com efeito, os valores próprios desta matriz são as raízes da equação

$$\lambda^3 + \frac{\lambda^2}{2} = 0,$$

donde

$$\lambda_1 = \lambda_2 = 0, \qquad \lambda_3 = -\frac{1}{2}\,.$$

Por conseguinte, o raio espectral de C_{GS} é

$$\rho(C_{GS}) = |\lambda_3| = \frac{1}{2}\,.$$

Logo, pelo Teorema 3.7, podemos confirmar que o método de Gauss-Seidel converge para a solução do sistema considerado, qualquer que seja a aproximação inicial considerada. ♦

3.4 Rapidez de convergência e análise do erro

Nos parágrafos precedentes estudámos condições que garantem a convergência dos métodos iterativos de Jacobi e de Gauss-Seidel. Atendendo aos resultados já obtidos, vamos compará-los quanto à rapidez de convergência.

Considerando qualquer norma vectorial V, e a norma matricial M a ela associada, podemos afirmar que, para qualquer método iterativo que verifique a igualdade (3.101), pág. 138, é satisfeita a desigualdade,

$$\|e^{(k+1)}\|_V \;\leq\; \|C\|_M \, \|e^{(k)}\|_V .$$

A rapidez de convergência depende das propriedades da matriz C e da aproximação inicial escolhida. Nalguns casos especiais pode acontecer que a solução exacta seja obtida após um número finito de iterações.

Na maioria dos casos com interesse prático, verifica-se que a ordem de convergência dos métodos aqui analisados é precisamente 1, ou seja, são de *convergência linear*.

Como sabemos, a rapidez de convergência de métodos da mesma ordem é caracterizada pelo factor assimpótico de convergência. Para avaliar esse factor, recorre-se frequentemente ao limite

$$c_1 \;=\; \lim_{k\to\infty} \frac{\|e^{(k+1)}\|_V}{\|e^{(k)}\|_V} \;. \tag{3.135}$$

A existência do limite c_1 depende das propriedades da matriz C e da norma V considerada. Além disso, para a mesma matriz C, o limite pode ter diferentes valores, conforme a aproximação inicial escolhida.

Pode mostrar-se que, se a matriz C tiver um *único* valor próprio $\lambda \in \mathbb{R}$, tal que $|\lambda| = \rho(C)$ (designado como *valor próprio dominante*), então para certas aproximações iniciais, o limite c_1 existe e verifica-se $c_1 = \rho(C)$. Logo, se o limite c_1 existir e o método iterativo convergir, tem-se $0 < c_1 < 1$ e este valor pode ser tomado como o factor assimptótico de convergência.

Assim, para valores de c_1 próximos de 0, teremos convergência rápida, enquanto que para valores de c_1 próximos de 1 teremos convergência lenta (isto é, são necessárias muitas iterações para atingir uma dada precisão).

Na prática o valor de c_1 não pode ser obtido directamente da fórmula (A.4), uma vez que os valores $\|e^{(k+1)}\|_V$ e $\|e^{(k)}\|_V$ não são, em geral, conhecidos para nenhuma iterada (visto que a solução $x = A^{-1}\, b$ é geralmente desconhecida). Por isso, recorre-se frequentemente às igualdades

$$\begin{aligned} x^{(k+1)} \,-\, x^{(k)} \;&=\; -e^{(k+1)} \,+\, e^{(k)} \;= \\ &=\; -C\, e^{(k)} \,+\, C\, e^{(k-1)} \;=\; C(x^{(k)} \,-\, x^{(k-1)}), \end{aligned} \tag{3.136}$$

donde se depreende que a diferença entre iteradas sucessivas varia com k do mesmo modo que o erro $e^{(k)}$ (ambas estas grandezas satisfazem uma relação do tipo (3.101), pág. 138). Logo, se o limite (A.4) existir, também existe o limite

$$c_1' = \lim_{k\to\infty} \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V} . \tag{3.137}$$

e os dois limites (c_1 e c_1') têm o mesmo valor, para certas aproximações iniciais.

Para se avaliar c_1, calcula-se para sucessivos valores de k, a razão

$$r^{(k)} = \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V},$$

até que o seu valor estabilize. O número assim obtido é tomado como uma estimativa de c_1.

Majorações de erro

Os valores do quociente $r^{(k)}$ também podem ser utilizados para obter estimativas do erro $e^{(k)}$.

Se considerarmos um valor c_2 tal que $r^{(k)} \leq c_2$, $\forall\, k > k_0$ (aqui k_0 representa a ordem a partir da qual o valor de $r^{(k)}$ estabiliza), podemos esperar que, para $k > k_0$, se verifique

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq c_2 \|x^{(k)} - x\|_V . \tag{3.138}$$

Da desigualdade triangular, temos

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + \|x^{(k+1)} - x\|_V . \tag{3.139}$$

De (3.139) e (3.138) resulta

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + c_2\|x^{(k)} - x\|_V,$$

donde

$$(1 - c_2) \|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V . \tag{3.140}$$

Uma vez que $c_2 < 1$, por construção, da desigualdade (3.140) obtém-se

$$\|e^{(k)}\|_V = \|x^{(k)} - x\|_V \leq \frac{\|x^{(k)} - x^{(k+1)}\|_V}{1 - c_2}. \tag{3.141}$$

Utilizando (3.138), de (3.141) obtém-se, sendo $c_2 < 1$,

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq \frac{c_2}{1 - c_2}\|x^{(k)} - x^{(k+1)}\|_V . \tag{3.142}$$

A desigualdade (3.142) permite-nos majorar o erro de uma dada iterada, bastando para tal conhecer a diferença entre as duas últimas iteradas e o valor de c_2.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k+1)} - x^{(k)}\|_2$	$r^{(k)}$
1	0.6	0.75	0.9	0.15	
2	0.625	0.85	0.875	0.106066	0.7071064
3	0.575	0.875	0.925	0.07500	0.7071066
4	0.5625	0.825	0.9375	0.05303	0.7071069
5	0.5875	0.8125	0.9125	0.03750	0.7071068
6	0.59375	0.8375	0.90625	0.02652	0.7071083
7	0.58125	0.84375	0.91875	0.01875	0.7071075
8	0.578125	0.83125	0.921875	0.01326	0.7071061
9	0.584375	0.828125	0.915625	0.00938	0.7071068

Tabela 3.1: Método de Jacobi para o Exemplo 3.13.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k+1)} - x^{(k)}\|_2$	$r^{(k)}$
1	0.6	0.8	0.9	0.141421	
2	0.6	0.85	0.925	0.055902	0.3952846
3	0.575	0.825	0.9125	0.037500	0.6708187
4	0.5875	0.8375	0.91875	0.018750	0.5
5	0.58125	0.83125	0.915625	0.009375	0.5

Tabela 3.2: Método de Gauss-Seidel para o Exemplo 3.13.

Exemplo 3.13. *Retomando o sistema linear do Exemplo 3.11, pág. 148, vamos efectuar uma análise do erro para os métodos de Jacobi e de Gauss-Seidel aplicados ao sistema.*

Partindo da aproximação inicial $x^{(0)} = (0.5, 0.8, 1.0)$, foram efectuadas iterações até satisfazer a condição

$$\|x^{(k)} - x^{(k+1)}\|_2 \leq 0.01 \ .$$

Em cada iteração foi avaliada a norma $\|x^{(k)} - x^{(k+1)}\|_2$, e a partir da $2^{\underline{a}}$ iteração, a razão $r^{(k)}$ correspondente. Os resultados obtidos para o método de Jacobi são dados na Tabela 3.1, enquanto os resultados obtidos para o método de Gauss-Seidel se encontram na Tabela 3.2.

Verifica-se numericamente que os valores de $r^{(k)}$ tendem para $c_1 = 0.7071$, no caso do método de Jacobi, e para $c_1 = 0.5$, no método de Gauss-Seidel. Estes valores coincidem com os raios espectrais das matrizes C_J e C_{gs}, respectivamente (ver Exemplo 3.11, pág. 148, e Exemplo 3.12, pág. 151).

Com base nestes valores, podemos obter estimativas do erro para cada um dos métodos. Para o método de Jacobi, de acordo om a fórmula (3.141), conside-

rando $c_2 = 0.70711$, temos

$$\|e^{(9)}\|_2 \leq \frac{c_2}{1 - c_2}\|x^{(9)} - x^{(8)}\|_2 \leq 0.0242 \ .$$

No caso do método de Gauss-Seidel, tomando $c_2 = 0.5$, temos

$$\|e^{(5)}\|_2 \leq \frac{c_2}{1 - c_2}\|x^{(5)} - x^{(4)}\|_2 \leq 0.01 \ .$$

♦

Comparação dos métodos de Jacobi e de Gauss-Seidel

No exemplo anterior constatámos que o método de Gauss-Seidel converge mais rapidamente que o de Jacobi, o que resulta de o raio espectral da matriz C_{GS} ser inferior ao da matriz C_J.

A fim de compararmos o método de Gauss-Seidel com o de Jacobi, quanto à rapidez de convergência, consideremos o caso em que a matriz A do sistema possui diagonal estritamente dominante por linhas. De acordo com o Teoremas 3.9, pág. 147, e Teorema 3.11, pág. 150, ambos os métodos convergem para a solução exacta, qualquer que seja a aproximação inicial escolhida.

Além disso, para o método de Jacobi é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \mu^k \, \|e^{(0)}\|_\infty, \qquad k = 1, 2, \ldots, \tag{3.143}$$

onde $\mu = \|C_J\|_\infty$. Recordando a forma da matriz C_J, dada por (3.95), pág. 136, e as definições das grandezas α_i e β_i, dadas por (3.127), pág. 150, podemos concluir que

$$\mu \ = ||C_J||_\infty = \max_{i=1,\ldots,n} (\alpha_i + \beta_i) \ . \tag{3.144}$$

Por outro lado, para o método de Gauss-Seidel, segundo o Teorema 3.11, é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \eta^k \, \|e^{(0)}\|_\infty, \ k = 1, 2, \ldots, \quad \text{com} \quad \eta = \max_{i=1:n} \frac{\beta_i}{1 - \alpha_i}, \tag{3.145}$$

desde que $\eta < 1$. Para estabelecer uma relação entre a rapidez de convergência dos dois métodos, basta-nos portanto comparar o parâmetro μ da fórmula (3.143) com o parâmetro η da fórmula (3.145).

Exemplo 3.14. *Consideremos o sistema $A\,x = b$, onde A é uma matriz tridiagonal*[21]*, de ordem $n \geq 2$, da forma*

$$A = \begin{bmatrix} 5 & 2 & 0 & \ldots & 0 \\ 2 & 5 & 2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ 0 & \ldots & 2 & 5 & 2 \\ 0 & \ldots & 0 & 2 & 5 \end{bmatrix}.$$

[21] Trata-se de uma matriz definida positiva. Ver adiante parágrafo 3.6, pág. 163.

Compare-se a rapidez de convergência do método de Jacobi e do método de Gauss-Seidel.

A matriz A possui a diagonal estritamente dominante por linhas, pelo que tanto o método de Gauss-Seidel como o de Jacobi convergem, qualquer que seja a aproximação inicial.

Atendendo às fórmulas (3.127), pág. 150, temos

$$\begin{array}{l} \alpha_1 = 0, \quad \alpha_i = 2/5, \quad \text{para} \quad i = 2 : n \\ \beta_i = 2/5, \quad \text{para} \quad i = 1 : (n-1), \quad \text{e} \quad \beta_n = 0 \ . \end{array}$$

De (3.144) e (3.128), resulta

$$\mu = 4/5, \ \ \eta = 2/3 \ .$$

Assim, neste exemplo verifica-se a desigualdade $\eta < \mu$. Por conseguinte, é de esperar que aqui o método de Gauss-Seidel convirja mais rapidamente que o de Jacobi.

Note-se porém que esta comparação entre os dois métodos só é válida para matrizes com a diagonal estritamente dominante por linhas. No caso geral nem sempre o método de Gauss-Seidel é mais rápido que o de Jacobi, havendo mesmo casos particulares em que o segundo é convergente e o primeiro não (ver adiante Exemplo 3.15, pág. 160). ⧫

Estabilidade numérica

É de realçar que os métodos iterativos para sistemas lineares, uma vez escolhida uma qualquer aproximação inicial, quando convergem são *estáveis* (ver Definição 3.9, pág. 127). Ou seja, partindo de dois vectores iniciais próximos, ξ_0 e η_0, obtêm-se sempre duas sucessões $(x_n)_{n\geq n_0}$ e $(y_n)_{n\geq n_0}$ igualmente próximas, convergindo para o mesmo vector x (solução exacta).

Esta propriedade, dita de *estabilidade numérica* é de grande importância prática, uma vez que no cálculo numérico são inevitáveis os erros de arredondamento, os quais se podem propagar ao longo de sucessivas operações, conduzindo a erros muito grandes no resultado final. Esta situação verifica-se, por exemplo, na resolução de sistemas lineares por métodos directos, mesmo que eles sejam bem condicionados.

Os métodos iterativos, desde que sejam aplicados a sistemas bem condicionados, são sempre estáveis, ou seja, quando se usam estes métodos não há perigo de os erros de arredondamento cometidos nos cálculos poderem resultar em erros significativos no resultado final. Isto representa, portanto, uma importante vantagem dos métodos iterativos sobre os directos, sobretudo quando se trata de resolver sistemas de grandes dimensões.

De facto, um algoritmo iterativo para a resolução de um sistema linear $A\,x = b$, por comparação com um método directo, oferece desde logo a vantagem de não modificar a matriz A ao longo do processo. Assim, mesmo que o algoritmo iterativo necessite de muitas iterações para aproximar a solução do sistema dentro de uma tolerância de erro predefinida, o problema da acumulação de erros de arredondamento ao longo do processo é em geral irrelevante por comparação com o que acontece nos métodos directos, nos quais a matriz A é modificada em cada passo. Nos métodos directos, a acumulação de erros de arredondamento pode ser muito grande, conforme se referiu no parágrafo 3.2.3, pág. 110.

3.5 Método das relaxações sucessivas (SOR) *

Neste parágrafo estudaremos uma generalização do método de Gauss-Seidel, muito utilizada no cálculo numérico, conhecida como *método das relaxações sucessivas* ou *método SOR* (acrónimo de "successive overrelaxation").

A interpretação geométrica do método é simples (ver Figura 3.2).

A partir de uma aproximação $x^{(k)}$ da solução do sistema $A\,x = b$, aplica-se o método de Gauss-Seidel para calcular outra aproximação $z^{(k+1)}$. O objectivo é escolher uma nova aproximação $x^{(k+1)}$, tal que o vector $x^{(k+1)} - x^{(k)}$ seja colinear com o vector $z^{(k+1)} - x^{(k)}$, de modo que o ponto $x^{(k+1)}$ esteja mais próximo da solução $A^{-1}\,b$ do que estava o ponto de partida $x^{(k)}$ e o ponto $z^{(k+1)}$, obtido pelo método de Gauss.

Evidentemente que a posição do ponto $x^{(k+1)}$ depende do valor atribuído ao parâmetro de controle ω.

Estamos por conseguinte a falar de uma família de métodos dependente de um parâmetro ω, cuja matriz de iteração (comparar com a expressão (3.151), pág. 159) pode ser escrita na forma,

$$C_\omega = -M_\omega^{-1} N_\omega, \tag{3.146}$$

onde

$$M_\omega = L + \frac{1}{\omega}\,D, \qquad N_\omega = U + (1 - \frac{1}{\omega})\,D, \tag{3.147}$$

sendo as matrizes L,D e U definidas como no caso dos métodos de Jacobi e de Gauss-Seidel.

É fácil verificar que, no caso de $\omega = 1$, se obtém $M_1 = L + D$ e $N_1 = U$, pelo que $C_1 = -(L + D)^{-1}U$, ou seja, neste caso o método SOR reduz-se ao método de Gauss-Seidel.

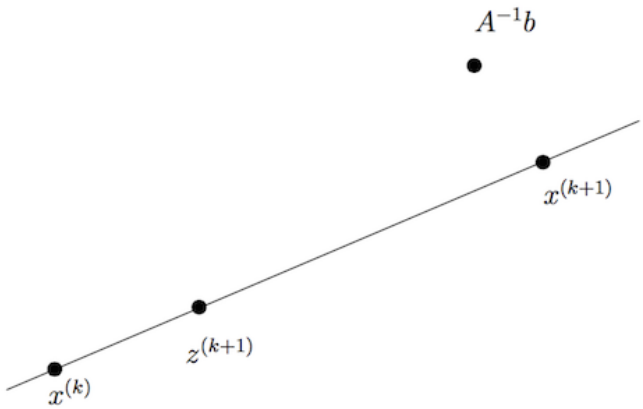


Figura 3.2: Geometria do método SOR. Se $\omega < 1$, o ponto $x^{(k+1)}$ pertence ao segmento $[x^{(k)}, z^{(k+1)}]$; se $\omega > 1$, o mesmo ponto ocupa uma posição como a figurada.

Atendendo a que $x^{(k+1)} = (1-\omega)\, x^{(k)} + \omega\, z^{(k+1)}$, as fórmulas computacionais do método SOR escrevem-se,

$$x^{(k+1)} = \omega z^{(k+1)} + (1-\omega) x^{(k)}, \tag{3.148}$$

onde

$$z^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}}{a_{ii}}, \qquad i = 1:n, \tag{3.149}$$

é a $(k+1)$-ésima iterada do método de Gauss-Seidel. Assim, podemos dizer que cada iterada do método SOR é uma média ponderada entre a nova iterada (obtida pelo método de Gauss-Seidel) e a iterada anterior, sendo ω o peso da nova iterada.

Ao introduzir o parâmetro ω ficamos por vezes habilitados a melhorar a convergência do método de Gauss-Seidel. Isso consegue-se estudando o raio espectral da matiz C_ω como função de ω, de modo a escolher um valor de ω que minimize esse raio espectral, ou experimentalmente testando diferentes valores para o parâmetro ω.

3.5.1 Condição necessária de convergência

O resultado a seguir mostra-nos que o parâmetro ω do método deverá ser escolhido no intervalo $(0, 2)$.

Teorema 3.12. Se o método SOR converge para a solução de um sistema linear $A\, x = b$, então

$$0 < \omega < 2\ .$$

Demonstração. Atendendo às expressões (3.148) e (3.149), tem-se

$$a_i\, x_i^{(k+1)} + \omega \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} = (1-\omega)\, a_{ii}\, x_i^{(k)} - \omega \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} + \omega\, b_i, \quad i = 1:n\ .$$

A expressão anterior pode ser rescrita em termos das matrizes D, L e U, como

$$(D + \omega\, L)\, x^{(k+1)} = ((1-w)\, D - \omega\, U)\ x^{(k)} + \omega\, b\ . \tag{3.150}$$

As matrizes que entram na expressão (3.150) têm uma forma particular:

$$D + \omega\, L = D\,(I + \omega\, D^{-1}\, L) = D\,(I + \omega\, E),$$

onde $E = D^{-1}\, L$ é uma matriz triangular inferior em que a diagonal principal é nula. Pelo seu lado, a matriz

$$(1-\omega)\, D - \omega\, U = D\ ((1-\omega)\, I - \omega\, F)\,,$$

onde $F = D^{-1}\, U$ é uma matriz triangular superior, com a diagonal principal nula. Levando em consideração as expressões anteriores (3.150), resulta

$$x^{(k+1)} = (I + \omega\, E)^{-1}\ ((1-\omega)\, I - w\, F)\ x^{(k)} + \omega\, (D + \omega\, L)^{-1}\, b\ .$$

Por conseguinte, a matriz de iteração do método é da forma

$$C_\omega = (I + \omega\, E)^{-1}\ ((1-\omega)\, I - \omega\, F)\ . \tag{3.151}$$

O primeiro factor da matriz C_ω é uma matriz triangular inferior com diagonal unitária, pelo que o respectivo determinante vale 1. O segundo factor de C_ω é uma matriz triangular superior cuja diagonal principal é constituída por entradas todas iguais a $1 - w$. Por conseguinte,

$$|\det(C_\omega)| = |\det((1-\omega)\, I - \omega\, F)| = |1-\omega|^n.$$

Sendo $\lambda_1, \lambda_2, \ldots, \lambda_n$ o espectro da matriz C_ω, tem-se que

$$|\det(C_\omega)| = |\lambda_1| \times |\lambda_2| \times \ldots \times |\lambda_n|\ .$$

Logo,

$$|1-\omega|^n = |\lambda_1| \times |\lambda_2| \times \ldots \times |\lambda_n| \leq \rho(C_\omega)^n,$$

ou, equivalentemente,

$$|1-\omega| \leq \rho(C_\omega)\ .$$

Uma vez que o método SOR é, por hipótese, convergente para a solução de $A\, x = b$, necessariamente $\rho(C_\omega) < 1$, ou seja,

$$|1-w| < 1 \Longleftrightarrow 0 < \omega < 2\ .$$

□

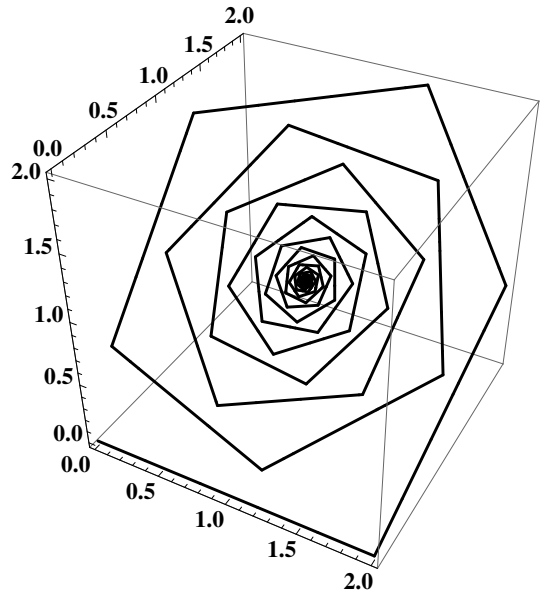


Figura 3.3: Partindo de $x^{(0)} = (0, 0, 0)$, efectuaram-se 150 iterações. O método de Jacobi converge muito lentamente (ver Exemplo 3.15).

Se no método SOR fixarmos $0 < \omega < 1$, dizemos que ω é um parâmetro de *sub–relaxação*. Se $1 < \omega < 2$, dizemos que ω é parâmetro de *sobre-relaxação*.

No exemplo a seguir é dado um sistema linear para o qual o método de Jacobi é convergente, embora a convergência seja muito lenta. Acontece que o método de Gauss-Seidel não converge. Mostramos que é possível escolher um parâmetro de sub-relaxação para o qual o método SOR é convergente e de convergência mais rápida do que o método de Jacobi.

Exemplo 3.15. *Considere o sistema linear* $A\,x = b$,

$$\begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix},$$

de solução $x = (1, 1, 1)$. *Mostremos que:*
(a) O método de Jacobi converge e que a convergência é lenta.

(b) O método de Gauss-Seidel não é convergente.

(c) Escolhido um parâmetro de sub-relaxação o método SOR é convergente, mas não é convergente se usarmos sobre-relaxação. Escolhido o parâmetro de sub--relaxação óptimo, $\omega_{opt} = 2/3$, *o método converge mais rapidamente do que o método de Jacobi.*

(a) A matriz de iteração para o método de Jacobi é

$$C_J = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 1/3 & 2/3 & 0 \end{bmatrix}.$$

Dado que $||C_J||_\infty = 1$, sabemos que $\rho(C_J) \leq ||C_J||_\infty \leq 1$.

Vejamos que a matriz de iteração não pode possuir nenhum valor próprio de módulo unitário e, consequentemente, o seu raio espectral é necessariamente inferior à unidade. A equação característica da matriz de iteração, $det(C_J - \lambda\, I) = 0$, é

$$\lambda^3 + \frac{1}{3}\,(2+\lambda) = 0\ . \tag{3.152}$$

Sabemos que nenhum valor próprio da matriz de iteração possui módulo superior a 1. Suponhamos que existe $\lambda \in \mathbb{C}$, tal que $|\lambda| = 1$. Iremos concluir que esta hipótese não se verifica, pelo que necessariamente todos os valores próprios possuem módulo inferior à unidade.

De (3.152), resulta

$$|\lambda^3| = \frac{1}{3}\,|2+\lambda|, \quad \text{donde} \quad 3 = |2+\lambda|\ .$$

Ora, as condições $|\lambda| = 1$ e $|\lambda + 2| = 3$ são ambas satisfeitas apenas quando λ é real e $\lambda = 1$. Mas, $\lambda = 1$ não é raiz da equação (3.152). Conclui-se, portanto, que $\rho(C_J) < 1$, pelo que o método é convergente.

Pode verificar-se que o espectro aproximado de C_J é constituido por um número real e dois números complexos conjugados, isto é,

$$\{-0.747415, 0.373708 + 0.867355 \times i, 0.373708 - 0.867355 \times i\}.$$

Assim, $\rho(C_J) \simeq 0.944438$, o que indicia ser o método de convergência muito lenta.

Partindo do ponto $x^{(0)} = (0,0,0)$, mostra-se graficamente na Figura 3.3 a evolução do processo, após 150 iterações. As iteradas consecutivas são vértices da linha poligonal que aparece na figura. O gráfico sugere que o método de Jacobi converge para $x = (1,1,1)$, embora a convergência seja de facto muito lenta.

(b) Para o sistema dado, as fórmulas computacionais do método de Gauss-Seidel obtêm-se muito facilmente:

$$\begin{aligned} x_1^{(k+1)} &= 2 - x_3^{(k)} \\ x_2^{(k+1)} &= x_1^{(k+1)} = 2 - x_3^{(k)}, \qquad\qquad k = 0, 1, \ldots \\ x_3^{(k+1)} &= \frac{x_1^{(k+1)} + 2\, x_2^{(k+1)}}{3} = 2 - x_3^{(k)}. \end{aligned} \tag{3.153}$$

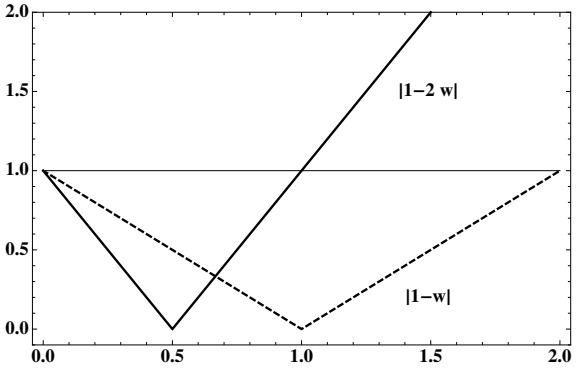


Figura 3.4: O método SOR não converge se $\omega \geq 1$.

Das fórmulas anteriores resulta imediatamente a respectiva matriz de iteração,

$$C_{GS} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Uma vez que a matriz anterior é triangular, o seu espectro obtém-se facilmente:

$$Sp(C_{GS}) = \{0, -1\}, \quad \text{logo} \quad \rho(C_{GS}) = 1 \ .$$

Por conseguinte, o método não converge. Com efeito, partindo de $x^{(0)} = (a, b, c) \in \mathbb{R}^3$, obtém-se

$$\begin{aligned} x^{(1)} &= (2-c, 2-c, 2-c) \\ x^{(2)} &= (c, c, c) \\ x^{(3)} &= (2-c, 2-c, 2-c) \\ &\vdots \end{aligned}$$

Exceptuando o caso $c = 1$, para o qual a sucessão de iteradas coincide com a solução $x = (1, 1, 1)$ do sistema, o método origina uma sucessão de vectores periódica, de período 2.

O comportamento observado não é de estranhar, porquanto qualquer vector da forma $v = (c, c, c)$ é vector próprio da matriz C_{GS} associado ao valor próprio $\lambda = -1$ (visto que $C_{GS}\, v = -v$). Consequentemente, ao partirmos de um vector de componentes iguais, como o vector v, a sucessão de iteradas é necessariamente periódica, obtendo-se: $-v$, v, $-v$, ...

É interessante relembrar aqui um comportamento análogo que pode ser observado no caso de funções iteradoras reais, geradoras de um processo iterativo a partir da equação de ponto fixo $x = g(x)$, para as quais um ponto z é ponto fixo *neutro* satisfazendo a condição $g'(z) = -1$ (ver secção 2.1.4, pág. 42).

(c) Uma vez que no método SOR se tem $x_{\text{novo}} = x + \omega\,(C_{GS}\, x + g_{GS} - x)$, a matriz de iteração do método é da forma

$$C_\omega = (1 - \omega)\, I + \omega\, C_{GS}.$$

Atendendo a (3.153), obtém-se

$$C_\omega = \begin{bmatrix} 1-\omega & 0 & 0 \\ 0 & 1-\omega & 0 \\ 0 & 0 & 1-\omega \end{bmatrix} + \begin{bmatrix} 0 & 0 & -\omega \\ 0 & 0 & -\omega \\ 0 & 0 & -\omega \end{bmatrix} = \begin{bmatrix} 1-\omega & 0 & 1-\omega \\ 0 & 1-\omega & -\omega \\ 0 & 0 & 1-2\,\omega \end{bmatrix}.$$

Assim,

$$Sp(C_\omega) = \{1-\omega, 1-2\,\omega\} \quad \text{e} \quad \rho(C_\omega) = max(|1-\omega|, |1-2\,\omega|)\ .$$

Na Figura 3.4 mostra-se os gráficos de $|1-\omega|$ e $|1-2\,\omega|$, no intervalo $(0,2)$. Uma vez que para $\omega \geq 1$ se tem $\rho(C_\omega) \geq 1$, concluimos imediatamente que se escolhermos um valor de sobre-relaxação o método SOR não converge. A convergência verifica-se se e só se $0 < \omega < 1$, ou seja, escolhendo um valor de sub-relaxação para o parâmetro ω.

A mesma figura sugere que existe um valor de ω óptimo, ω_{opt}, o qual se obtém minimizando o raio espectral da matriz. Ou seja, ω_{opt} satisfaz a equação

$$2\,\omega - 1 = 1 - \omega,$$

isto é,

$$\omega_{opt} = 2/3 \quad \Longrightarrow \quad \rho(C_{\omega_{opt}}) = 1/3\ .$$

Comparando com a alínea (a), conclui-se que o método SOR, para $\omega = 2/3$ converge mais rapidamente do que o método de Jacobi, pois $\rho(C_{\omega_{opt}}) < \rho(C_J)$. Relembre-se de que o método de Gauss-Seidel nem sequer é convergente.

Na Figura 3.5 mostram-se as primeiras cinco iteradas do método e os respectivos valores calculados, partindo de $x^{(0)} = (0,0,0)$. Note-se que o modelo de colinearidade adoptado (ver Figura 3.2, pág. 158) encontra aqui uma ilustração.

♦

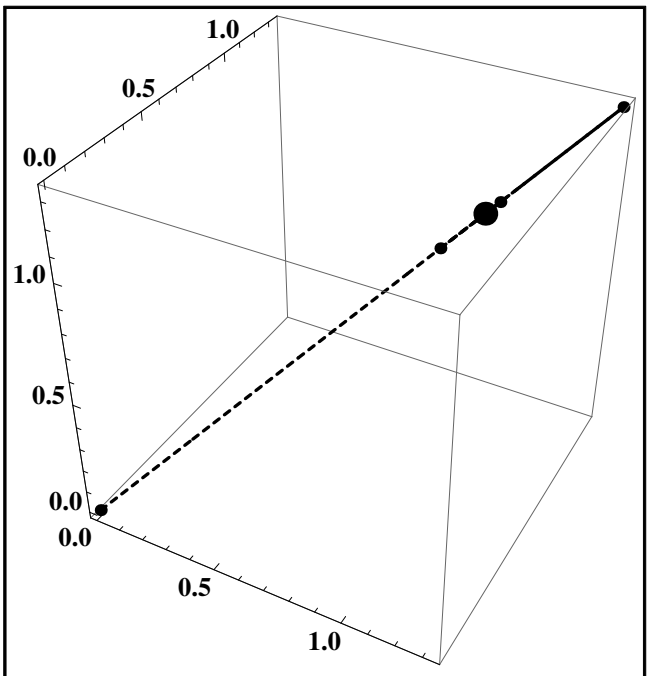
3.6 Matrizes simétricas definidas positivas

A classe das matrizes simétricas definidas positivas, a que fizemos referência na parágrafo, 3.2.7, pág. 122, ocorre com frequência nas aplicações. Em particular, os métodos SOR, do qual o método de Gauss-Seidel é um caso particular, são convergentes quando aplicados a sistemas de matriz definida positiva.

Comecemos por relembrar a definição de matriz definida positiva.

Definição 3.12. Uma matriz simétrica $A \in \mathbb{R}^{n\times n}$ diz-se definida positiva se e só se para qualquer vector $x \neq 0$ se verifica a desigualdade

$$x^T\, A\, x > 0, \quad \text{com} \quad x \in \mathbb{R}^n \quad \text{e} \quad x \neq 0.$$



0.	0.	0.
1.33333	1.33333	1.33333
0.888889	0.888889	0.888889
1.03704	1.03704	1.03704
0.987654	0.987654	0.987654
1.00412	1.00412	1.00412

Figura 3.5: Cinco iterações do método SOR com parâmetro óptimo $\omega_{opt} = 2/3$. O ponto negro de maior dimensão representa a solução do sistema.

Uma matriz simétrica definida positiva dispensa escolha parcial de pivot, porquanto as entradas da sua diagonal principal são mais 'pesadas' do que as entradas fora da diagonal. O Exemplo a seguir mostra que assim é para uma matriz simétrica 2×2, mas tal propriedade mantém-se para matrizes simétricas definidas positivas em geral.

Exemplo 3.16. *Seja*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

matriz (simétrica) definida positiva. Vejamos em que medida as entradas da diagonal são mais 'pesadas' do que as entradas fora da diagonal.

Fazendo sucessivamente $x = (1,0)^T$, $x = (0,1)^T$, $x = (1,1)^T$ e $x = (1,-1)^T$, resultam os seguintes valores de $x^T\, A\, x$:

$$\begin{aligned} &a_{11} > 0 \\ &a_{22} > 0 \\ &a_{11} + 2\, a_{12} + a_{22} > 0 \\ &a_{11} - 2\, a_{12} + a_{22} > 0 \ . \end{aligned}$$

Das duas últimas desigualdades, obtém-se

$$|a_{12}| \leq \frac{a_{11} + a_{22}}{2} \ .$$

A média anterior significa que as entradas da diagonal principal da matriz prevalecem, já que a sua grandeza é superior à grandeza da entrada fora da diagonal. ⧫.

Uma vez que nem sempre é fácil decidir da positividade de uma matriz simétrica a partir da Definição 3.12, é útil conhecer o critério dado a seguir.

Teorema 3.13. Uma matriz simétrica $A \in \mathbb{R}^{n\times n}$ é definida positiva se e só se todos os seus *menores principais*[a] são positivos, isto é,

$$det(A_k) > 0, \qquad k = 1 : n.$$

[a] Recorde-se que uma submatriz principal, A_k, obtém-se da matriz A suprimindo as últimas $n-k$ linhas e colunas de A.

Demonstração. Ver, por exemplo, [35], pág. 58-59. □

É oportuno relembrar aqui uma aplicação do Teorema 3.13, dada no Exemplo 3.91, pág. 133.

O Teorema 3.14 a seguir garante convergência do método SOR para sistemas lineares cuja matriz pertence à classe das matrizes simétricas reais *definidas positivas*.

Teorema 3.14. Sendo A uma matriz (simétrica) real definida positiva, o método SOR, com $0 < w < 2$, converge para a solução de um sistema $A\,x = b$ dado.

Demonstração. Ver [16], pág. 512. □

Do Teorema 3.14 resulta, em particular, que o método de Gauss-Seidel ($\omega = 1$) é sempre convergente quando aplicado a sistemas de matriz simétrica definida positiva.

3.6.1 Sistemas de grandes dimensões

Sistemas de grandes dimensões ocorrem naturalmente quando se aproxima a solução (contínua) de certos tipos de equações diferenciais por um vector que resulta da chamada *discretização* de um problema diferencial, tal como é sugerido no Exemplo 3.17, onde se pretende aproximar a solução de um problema que recebe a designação de *problema de valores de fronteira*. Problemas desta natureza estão fora do âmbito deste curso. O exemplo a seguir serve apenas para ilustrarmos um caso onde é necessário resolver sistemas que podem ser de dimensão muito elevada.

Exemplo 3.17. *Considere o problema*

$$\begin{cases} y''(t) = 2\,t, & 0 \leq t \leq 1 \\ y(0) = \alpha, & y(1) = \beta, \end{cases} \tag{3.154}$$

onde as constantes α e β são dados. Pode verificar-se que a função polinomial

$$y(t) = \alpha + \frac{t}{3}\,(t^2 + 3\,(\beta - \alpha) - 1),$$

é solução de (3.154)*. Efectuando uma discretização adequada, aproximamos a solução anterior resolvendo um sistema linear.*

Fixado um número natural N, comecemos por subdividir o intervalo $[a, b] = [0, 1]$ em $N + 1$ partes, e considerem-se os $N + 2$ pontos equidistantes do intervalo,

$$t_i = i\,h, \qquad i = 0 : (N + 1),$$

onde $h = 1/(N + 1)$ é o espaçamento entre pontos consecutivos.

Conhecemos os valores da solução do problema nos extremos do intervalo. Sejam $y_0 = \alpha$ e $y_{N+1} = \beta$. Designemos por $y_1, y_2, \ldots, y_N$, as aproximações da solução $y(t)$ em cada um dos pontos interiores t_i do intervalo em causa.

Partindo dos desenvolvimentos de Taylor da solução $y(t)$,

$$\begin{aligned} y(t + h) &= y(t) + h\,y'(t) + h^2/2\,y''(t) + h^3/3!\,y'''(t) + \mathcal{O}(h^4) \\ y(t - h) &= y(t) - h\,y'(t) + h^2/2\,y''(t) - h^3/3!\,y'''(t) + \mathcal{O}(h^4), \end{aligned}$$

somando membro a membro as igualdades anteriores e após isolarmos $y''(t)$, facilmente se conclui que

$$y''(t) = \frac{y(t-h) - 2\,y(t) + y(t+h)}{h^2} + \mathcal{O}(h^2).$$

Assim, a segunda derivada $y''(t)$ pode ser *aproximada*, em cada ponto t_i do intervalo $[0,1]$, através da expressão

$$y_i'' = \frac{y_{i-1} - 2\,y_i + y_{i+1}}{h^2}. \tag{3.155}$$

Ao substituirmos em (3.154) a derivada y'', em cada ponto t_i do intervalo, pela aproximação dada por (3.155), obtemos as seguintes N equações lineares:

$$\begin{array}{ll}
i = 1 \rightarrow & y_0 - 2\,y_1 + y_2 = 2\,t_1\,h^2 = 2\,h^3 \Longleftrightarrow -2\,y_1 + y_2 = 2\,h^3 - \alpha \\
i = 2 \rightarrow & y_1 - 2\,y_2 + y_3 = 2\,t_2\,h^2 = 4\,h^3 \\
& \quad \vdots \\
i = N-1 \rightarrow & y_{N-2} - 2\,y_{N-1} + y_N = 2\,t_{N-1}\,h^2 = 2\,(N-1)\,h^3 \\
i = N \rightarrow & y_{N-1} - 2\,y_N + y_{N+1} = 2\,t_N\,h^2 = 2\,N\,h^3 \\
& \qquad\qquad \Longleftrightarrow y_{N-1} - 2\,y_N = 2\,N\,h^3 - \beta.
\end{array}$$

Assim, o sistema a resolver é da forma

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & 0 & \ldots & -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = h^3 \begin{bmatrix} \alpha/h^3 - 2 \\ -4 \\ \vdots \\ -2\,(N-1) \\ \beta/h^3 - 2\,N \end{bmatrix}.$$

A matriz do sistema é tridiagonal simétrica. Trata-se de matriz definida positiva, conforme poderá comprovar aplicando o critério dos menores enunciado no Teorema 3.13 (pág. 165), pelo que tanto os métodos de Gauss-Seidel como SOR são aplicáveis para determinar aproximações da solução do sistema.

A estrutura tridiagonal simétrica da matriz anterior sugere que se construa um processo directo para determinar a solução exacta. No entanto, pode ser mais interessante usar um dos métodos iterativos referidos, caso a dimensão da matriz seja grande.

Por exemplo, para $h = 10^{-5}$, o número N de equações em jogo é $N = 10^5$. Trata-se de um sistema *esparso* (i.e, com uma grande quantidade de entradas nulas), a partir do qual facilmente se obtêm as fórmulas computacionais para os métodos de Gauss-Seidel ou SOR.

Convida-se o leitor a fazer uma simulação numérica, de modo a comparar os valores calculados para y_i, com os valores exactos $y(t_i)$, para $h = 10^{-j}$, com $j = 1:5$. ♦

3.7 Métodos iterativos para sistemas não lineares

Consideremos um sistema de n equações não lineares, da forma

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \qquad \ldots \\ F_n(x_1, x_2, \ldots, x_n) = 0 \ . \end{cases} \tag{3.156}$$

onde cada uma das funções F_i é uma função real de n variáveis reais. Este sistema pode ser escrito na forma vectorial

$$F(x) = 0,$$

onde $F = (F_1, F_2, \ldots, F_n)$ e $x = (x_1, x_2, \ldots, x_n)$. O ponto $z \in \mathbb{R}^n$ diz-se solução (ou raiz) do sistema (3.156) se $F(z) = 0$.

O problema da determinação das raízes de um sistema não linear é complexo e em geral mais difícil do que no caso de sistemas lineares. Em primeiro lugar, não há nenhum critério simples que nos permita verificar se o sistema (3.156) tem ou não solução. No caso de existirem várias soluções, não é fácil isolar cada uma, isto é, definir um conjunto em $\mathbb{R}^n$ que contenha essa raiz e não contenha outras.

Uma das abordagens para localização de raízes de um sistema não linear é baseada no teorema do ponto fixo (que pode ser reformulado para funções de $\mathbb{R}^n$ em $\mathbb{R}^n$) e que discutiremos adiante. O mesmo teorema permite-nos definir um método iterativo (método do ponto fixo em $\mathbb{R}^n$) para aproximar as raízes do sistema. Finalmente, veremos que o método de Newton (estudado na Secção 2.3, pág. 65, no caso de uma equação) pode ser generalizado para sistemas de n equações.

3.7.1 Método do ponto fixo em $\mathbb{R}^n$ *

A fim de investigarmos condições que garantem a convergência do método do ponto fixo em $\mathbb{R}^n$, vamos formular uma generalização do *teorema do ponto fixo*, estudado no parágrafo 2.1.4, pág. 42. Com esse objectivo, necessitamos de introduzir algumas definições.

Definição 3.13. Seja E um espaço normado, $X \subset E$ e G uma função de X em E. A função G diz-se lipschitziana em X, se existir uma constante q, tal que

$$\|G(x_1) - G(x_2)\| \leq q \, \|x_1 - x_2\|, \ \ \forall \, x_1, x_2 \in X \ . \tag{3.157}$$

Ao ínfimo de todas as constantes q, para as quais a desigualdade (3.157) é satisfeita, chama-se *constante de Lipschitz*[22] de G em X e representa-se por $L_{G,X}$.

[22]Rudolf Otto Sigismund Lipschitz, 1832 – 1903, matemático alemão.

Definição 3.14. Diz-se que G é uma contracção (ou uma função contractiva) em X se G for lipschitziana em X, e

$$L_{G,X} < 1.$$

Exemplo 3.18. *Seja $E = \mathbb{R}$ e $G(x) = x^2$. Indaguemos para que valores de r a função G é contractiva em $X = [-r, r]$.*

Temos

$$|G(x_1) - G(x_2)| = |x_1^2 - x_2^2| = |x_1 - x_2|\,|x_1 + x_2| \,. \tag{3.158}$$

Se x_1 e x_2 pertencerem a X, podemos escrever

$$|x_1 + x_2| \leq r + r = 2\,r \,. \tag{3.159}$$

Substituindo (3.159) em (3.158), obtém-se

$$|G(x_1) - G(x_2)| \leq 2\,r\,|x_1 - x_2|,$$

donde se conclui que G é lipschitziana em X, com a constante $L_{G,X} = 2\,r$.

Por conseguinte, se $r < 1/2$, podemos afirmar que G é contractiva em X. ⧫

No caso de função de uma variável real, a condição de contractividade pode ser expressa noutros termos, tornando-se mais fácil a sua verificação.

Teorema 3.15. Seja G uma função real com domínio em $X = [a, b]$ e suponhamos que $G \in C^1([a, b])$. A função G é contractiva em X se e só se

$$\max_{x\in[a,b]} |G'(x)| < 1 \,. \tag{3.160}$$

Demonstração. Pelo teorema de Lagrange, quaisquer que sejam x_1 e x_2, pertencentes a $[a, b]$, existe $\xi \in (x_1, x_2)$, tal que

$$|G(x_1) - G(x_2)| = |G'(\xi)||x_1 - x_2| \,.$$

Assim, podemos afirmar que a constante de Lipschitz de G é

$$L_G = \max_{x\in[a,b]} |G'(x)| < 1,$$

donde se conclui que G é contractiva em $[a, b]$.

Para mostrarmos o recíproco, suponha-se que existe y em $[a, b]$, tal que $|G'(y)| \geq 1$. Sabemos que, pelo teorema de Lagrange, para qualquer $h > 0$, existe $\theta \in (0, 1)$, tal que

$$|G(y + h) - G(y)| = |G'(y + \theta h)|\,h \,. \tag{3.161}$$

Visto que G' é contínua em $[a,b]$, para qualquer $\rho < 1$, existe h_0 tal que

$$|G'(y+\theta h_0)| > \rho.$$

Escrevendo a desigualdade (3.161) com $h = h_0$, obtém-se

$$|G(y+h_0) \; - \; G(y)| \; = \; |G'(y+\theta h_0)|\, h_0 > \rho\, h_0 \; . \tag{3.162}$$

A desigualdade (3.162) implica que G não é contractiva em $[a,b]$, ficando assim demonstrado o teorema. □

O Teorema 3.15 permite-nos substituir a condição de contractividade pela condição (3.160), quando se consideram funções de uma variável. Foi isso precisamente o que fizemos na Secção 2.1.6, pág. 47.

Tal é generalizável ao caso em que G é uma função de $\mathbb{R}^n$ em $\mathbb{R}^n$, com derivadas parciais contínuas, onde a contractividade pode ser verificada através da matriz jacobiana de G.

Definição 3.15. Seja G uma função vectorial, tal que

$$G(x) = (G_1(x), G_2(x), \ldots, G_n(x))\,,$$

onde G_i é uma função escalar com domínio em $X \subset \mathbb{R}^n$. Se existirem em X as derivadas parciais $\dfrac{\partial G_i}{\partial x_j}$, para $i,j = 1:n$, chama-se *jacobiana de* G (e representa-se por J_G), a matriz

$$J_G(x) \; = \; \begin{bmatrix} \dfrac{\partial G_1}{\partial x_1} & \cdots & \dfrac{\partial G_1}{\partial x_n} \\ \dfrac{\partial G_2}{\partial x_1} & \cdots & \dfrac{\partial G_2}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial G_n}{\partial x_1} & \cdots & \dfrac{\partial G_n}{\partial x_n} \end{bmatrix}. \tag{3.163}$$

Teorema 3.16. Seja X um conjunto convexo em $\mathbb{R}^n$ e $G : X \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ uma função de classe C^1 em X. Se

$$sup_{x\in X} \|J_G(x)\|_\infty \; < \; 1,$$

G é contractiva em X (segundo a norma do máximo).

Demonstração. Sejam x_1 e x_2 dois elementos de X. Segundo o teorema de Lagrange para funções de n variáveis, para cada função G_i, existe um ponto ξ_i, pertencente ao segmento (x_1, x_2), tal que

$$G_i(x_1) - G_i(x_2) \; = < \nabla G_i(\xi_i), x_1 - x_2 >, \tag{3.164}$$

onde $< \cdot, \cdot >$ designa o produto interno usual em $\mathbb{R}^n$, e ∇G_i designa o gradiente de G_i, ou seja,

$$\nabla G_i(x) \;=\; \left(\frac{\partial G_i}{\partial x_1}, \ldots, \frac{\partial G_i}{\partial x_n}\right), \quad i \in \{1, \ldots, n\}, \tag{3.165}$$

Note-se que todos os pontos ξ_i pertencem a X, uma vez que este conjunto é, por hipótese, convexo. De (3.164) e (3.165), obtém-se

$$\begin{aligned} |G_i(x_1) - G_i(x_2)| \;\leq\; & \max_{j=1,\ldots,n} |x_{1,j} - x_{2,j}| \sum_{j=1}^{n} \left|\frac{\partial G_i}{\partial x_j}(\xi_i)\right| = \\ & = \|\nabla G_i(\xi_i)\|_1 \, \|x_1 - x_2\|_\infty, \qquad i = 1 : n. \end{aligned} \tag{3.166}$$

Seja i' um índice para o qual se verifica

$$|G_{i'}(x_1) - G_{i'}(x_2)| \;=\; \|G(x_1) - G(x_2)\|_\infty \; .$$

No caso de $i = i'$, a desigualdade (3.166) toma o aspecto

$$\|G(x_1) - G(x_2)\|_\infty \;\leq\; \|\nabla G_{i'}(\xi_{i'})\|_1 \, \|x_1 - x_2\|_\infty \; . \tag{3.167}$$

Atendendo a que

$$\|\nabla G_{i'}(\xi_{i'})\|_1 \;\leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} \left|\frac{\partial G_i}{\partial x_j}(\xi_{i'})\right| = \|J_G(\xi_{i'})\|_\infty \;<\; 1, \tag{3.168}$$

de (3.167) resulta que G é contractiva em X, segundo a norma do máximo. □

Nalguns casos, pode ser mais cómodo considerar em $\mathbb{R}^n$ outras normas que não a do máximo, por exemplo, a norma $\|\cdot\|_1$. Por isso enunciamos a seguir um teorema análogo ao anterior.

Teorema 3.17. Seja X um conjunto convexo em $\mathbb{R}^n$, e $G : X \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ uma função de classe C^1 em $\mathbb{R}^n$. Se

$$sup_{x \in X} \|J_G(x)\|_1 \;<\; 1,$$

então a função G é contractiva em X (segundo a norma $\|\cdot\|_1$).

Demonstração. A prova pode ser obtida por argumentos semelhantes aos usados na demonstração do Teorema 3.16, pelo que é deixada como exercício. □

Estamos agora em condições de formular o teorema do ponto fixo, para espaços normados de dimensão finita, por exemplo, os espaços $\mathbb{R}^n$.

Teorema 3.18. (*Teorema do ponto fixo em R^n*).
Seja E um espaço normado de dimensão finita, e X um subconjunto fechado e convexo de E. Seja G uma função contractiva em X, tal que

$$G(X) \subset X \ .$$

São válidas as afirmações:
(1) G tem um único ponto fixo z em X.
(2) Se $(x^{(k)})_{k\geq 0}$ for a sucessão de termos em E tal que $x^{(0)} \in X$ e

$$x^{(k+1)} \ = \ G(x^{(k)}), \qquad \forall k \geq 0,$$

então $(x^{(k)})_{k\geq 0}$ converge para z.
(3) Se G satisfaz, em X, a desigualdade (3.157), pág. 168, com $q < 1$, então são válidas as desigualdades

$$\|x^{(n+1)} - z\| \ \leq \ q\,\|x^{(n)} - z\|, \qquad \forall n \geq 1 \ . \tag{3.169}$$

e

$$\|x^{(m)} - z\| \ \leq \ \frac{q^m}{1-q}\,\|x^{(1)} - x^{(0)}\|, \qquad \forall m \geq 1 \ . \tag{3.170}$$

Demonstração. Em primeiro lugar, note-se que se $x^{(0)} \in X$, então $x^{(k)} \in X$, $\forall\, k$, visto que $G(X) \subset X$.

Comecemos por provar que a sucessão referida no ponto (2) é convergente. Para tal, basta provar que se trata de uma sucessão de Cauchy.

Uma vez que G é contractiva em X, existe uma constante $q < 1$, tal que

$$\|G(x_1) - G(x_2)\| \ \leq \ q\,\|x_1 - x_2\|, \ \ \forall x_1, x_2 \in X \ .$$

Em particular, para dois termos consecutivos da sucessão considerada verifica-se,

$$\|x^{(k+1)} - x^{(k)}\| \ \leq \ q\,\|x^{(k)} - x^{(k-1)}\|, \ \ \forall k \in \mathbb{N} \ . \tag{3.171}$$

Sejam $x^{(m)}$ e $x^{(n)}$ dois termos quaisquer da sucessão, com $n > m$. Podemos escrever

$$\begin{aligned} \|x^{(n)} - x^{(m)}\| \ &= \|x^{(n)} - x^{(n-1)} \ + \ x^{(n-1)} - x^{(n-2)} + \ \ldots \ + x^{(m+1)} - x^{(m)}|| \leq \\ &\leq \|x^{(n)} - x^{(n-1)}\| \ + \ \|x^{(n-1)} - x^{(n-2)}\| \ + \ \cdots + \ \|x^{(m+1)} - x^{(m)}\| \ . \end{aligned} \tag{3.172}$$

Das desigualdades (3.171) e (3.172), obtém-se

$$\begin{aligned} \|x^{(n)} - x^{(m)}\| \ \leq \ (q^{n-m-1} + \cdots + q + 1)\,\|x^{(m+1)} - x^{(m)}\| \ \leq \\ q^m\,(q^{n-m-1} + \cdots + q + 1))\ \|x^{(1)} - x^{(0)}\| \ . \end{aligned} \tag{3.173}$$

A adição que figura no segundo membro de (3.173) é a soma de uma progressão geométrica de razão q. Como $q < 1$, é válida a desigualdade

$$q^m \sum_{k=0}^{n-m-1} q^k \,<\, \frac{q^m}{1-q}, \;\; \forall n \in \mathbb{N} \; . \tag{3.174}$$

Substituindo (3.174) em (3.173), obtém-se

$$\|x^{(m)} - x^{(n)}\| < \frac{q^m}{1-q} \, \|x^{(1)} - x^{(0)}\|, \;\; \forall n > m \; . \tag{3.175}$$

Da desigualdade (3.175) resulta que $\forall \epsilon > 0$, existe $n_0 \in \mathbb{N}$ tal que

$$\|x^{(m)} - x^{(n)}\| \,<\, \epsilon, \qquad \forall m, n > n_0 \; . \tag{3.176}$$

Assim, a sucessão considerada é uma sucessão de Cauchy, logo convergente. Representemos por z o seu limite. Uma vez que X é fechado, $z \in X$.

Provemos agora que z é um ponto fixo de G. Utilizando o facto de G ser contractiva, podemos escrever

$$\|x^{(m+1)} - G(z)\| \,=\, \|G(x^{(m)}) - G(z)\| \,\leq\, q\|x^{(m)} - z\|, \qquad \forall\, m \; . \tag{3.177}$$

Logo $\|x^{(m+1)} - G(z)\| \to 0$, ou seja, $x^{(m)} \to G(z)$, quando $m \to \infty$. Por conseguinte, $G(z) = z$. Fica assim demonstrado o item (2) do teorema.

A desigualdade (3.169), por sua vez, resulta de (3.177). Quanto à desigualdade (3.170), ela obtém-se de (3.175), se fizermos n tender para infinito.

Resta-nos provar que z é o único ponto fixo de G em X.

Suponhamos que existem dois pontos fixos de G em X, e representemo-los por z e z'. Uma vez que G é contractiva, temos

$$\|G(z') - G(z)\| \,=\, \|z' - z\| \,\leq\, q\|z' - z\|,$$

donde

$$\|z' - z\| \, (1 \,-\, q) \,\leq\, 0 \; . \tag{3.178}$$

Dado que $1 - q > 0$, de (3.178) resulta que $z' = z$. □

Exemplo 3.19. *Consideremos o sistema de duas equações,*

$$\left\{ \begin{array}{rcrcl} 3\,x_1 & + & x_2^2 & = & 0 \\ x_1^2 & + & 3\,x_2 & = & 1 \; . \end{array} \right. \tag{3.179}$$

Vamos utilizar o teorema do ponto fixo para provar que este sistema tem uma única raiz no conjunto

$$X \,=\, \{(x_1, x_2) \in \mathbb{R}^2 : \quad -1/3 \leq x_1 \leq 0 \quad e \quad 0 \leq x_2 \leq 1/3\} \; .$$

O sistema (3.179) pode ser reescrito na forma $x = G(x)$, onde

$$\begin{aligned} G_1(x_1, x_2) &= -\frac{x_2^2}{3} \\ G_2(x_1, x_2) &= \frac{1 - x_1^2}{3} \ . \end{aligned} \tag{3.180}$$

Verifiquemos se a função $G = (G_1, G_2)$, definida por (3.180), satisfaz as condições do teorema do ponto fixo em X.

Em primeiro lugar, constata-se que o conjunto X é um quadrado, contendo a sua fronteira, pelo que é convexo e fechado. Além disso, as derivadas parciais de G_1 e G_2 são contínuas em X. A matriz jacobiana de G é

$$J_G(x_1, x_2) = \begin{bmatrix} 0 & -\frac{2\,x_2}{3} \\ -\frac{2\,x_1}{3} & 0 \end{bmatrix} . \tag{3.181}$$

Assim,

$$\|J_G(x_1, x_2)\|_\infty = \max_{(x_1,x_2)\in X} \left(\frac{2\,|x_2|}{3}, \frac{2\,|x_1|}{3} \right),$$

e portanto

$$\|J_G(x_1, x_2)\|_\infty \leq \frac{2}{9} < 1, \qquad \forall (x_1, x_2) \in X \ .$$

Com base no Teorema 3.16, pág. 170, podemos afirmar que G é contractiva em X (segundo a norma do máximo), tendo por constante de contractividade $q = \frac{2}{9}$.

Para se aplicar o teorema do ponto fixo, precisamos também de verificar que $G(X) \subset X$.
Para $x = (x_1, x_2) \in X$, temos

$$\begin{aligned} G_1(x_1, x_2) &= -\frac{x_2^2}{3} &&\in [-1/3, 0] \\ G_2(x_1, x_2) &= \frac{1 - x_1^2}{3} &&\in [0, 1/3] \ . \end{aligned} \tag{3.182}$$

Por conseguinte, $(G_1(x_1, x_2), G_2(x_1, x_2)) \in X$, de onde se conclui que $G(X) \subset X$.

Visto que a função G satisfaz as condições do teorema do ponto fixo, podemos garantir que esta função tem um único ponto fixo em X, o qual, por construção, será a única raiz do sistema (3.179) em X.

Para aproximar a raiz considerada tomemos como aproximação inicial qualquer ponto do conjunto X, por exemplo, a origem das coordenadas $x^{(0)} = (0, 0)$.

Obtêm-se as seguintes aproximações:

$$x_1^{(1)} = G_1(0,0) = 0, \qquad x_2^{(1)} = G_2(0,0) = \frac{1}{3}$$

e

$$x_1^{(2)} = G_1(0,1/3) = -\frac{1}{27}, \qquad x_2^{(2)} = G_2(0,1/3) = \frac{1}{3}\ .$$

Como obter uma estimativa do erro da iterada $x^{(2)}$? De acordo com a desigualdade (3.170), pág.172, podemos escrever

$$\|x^{(2)} - z\|_\infty \ \leq\ \frac{q^2}{1-q}\,\|x^{(1)} - x^{(0)}\|_\infty,$$

onde $q = 2/9$. Neste caso, temos $\|x^{(1)} - x^{(0)}\|_\infty = 1/3$. Assim,

$$\|x^{(2)} - z\|_\infty \ \leq\ \frac{4}{63}\,\frac{1}{3} = \frac{4}{189}\ .$$

Esta última estimativa pode ser refinada se, em vez da desigualdade (3.170), aplicarmos a desigualdade

$$\|x^{(m+1)} - z\|_\infty \ \leq\ \frac{q}{1-q}\|x^{(m+1)} - x^{(m)}\|_\infty,$$

que também se pode deduzir facilmente. Obtém-se

$$\|x^{(2)} - z\|_\infty \ \leq\ \frac{q}{1-q}\|x^{(2)} - x^{(1)}\|_\infty = \frac{2}{189}. \tag{3.183}$$

♦

3.7.2 Método de Newton

Sabemos que no caso de funções de variável real, o método de Newton pode ser considerado como um caso particular do método do ponto fixo. Recorde-se que, dada uma função f (de uma variável real), a função iteradora do método de Newton tem a forma

$$g(x) = x - \frac{f(x)}{f'(x)}\ . \tag{3.184}$$

Ao considerar o sistema (3.156), pág. 168, em vez da função f temos uma função vectorial F (de n variáveis). Admitimos que todas as derivadas parciais de F existem e são contínuas num certo conjunto D, onde se procura a raiz do sistema e que a matriz jacobiana de F,

$$J_F(x) = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1} & \dfrac{\partial F_1}{\partial x_2} & \cdots & \dfrac{\partial F_1}{\partial x_n} \\ \dfrac{\partial F_2}{\partial x_1} & \dfrac{\partial F_2}{\partial x_2} & \cdots & \dfrac{\partial F_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \dfrac{\partial F_n}{\partial x_2} & \cdots & \dfrac{\partial F_n}{\partial x_n} \end{bmatrix}$$

é invertível no domínio considerado.

Quando se generaliza o método de Newton para sistemas de equações, é natural substituir na fórmula (3.184) a expressão $1/f'(x)$ pela inversa da matriz jacobiana. Obtém-se assim formalmente a seguinte função iteradora para o método de Newton,

$$G(x) = x - J_F^{-1}(x)\, F(x).$$

Daqui resulta a fórmula iteradora do método de Newton para sistemas não lineares,

$$x^{(k+1)} = G(x^{(k)}) = x^{(k)} - J_F^{-1}(x^{(k)})\, F(x^{(k)}), \qquad k = 0, 1, 2 \ldots \tag{3.185}$$

onde $x^{(k)} \in \mathbb{R}^n$ representa a k-ésima iterada do método.

Fórmula computacional do método de Newton

A fórmula (3.185), embora definindo o método de Newton para sistemas (3.156), pág. 168, não é a que geralmente se aplica. Do ponto de vista computacional, não é vantajoso utilizar directamente esta fórmula, já que isso nos obrigaria, em cada iteração do método, a inverter uma matriz de ordem n (a jacobiana de F), o que seria muito dispendioso em termos de número de operações.

Comecemos por reescrever a fórmula iterativa na forma

$$x^{(k+1)} - x^{(k)} = -J_F^{-1}(x^{(k)})\, F(x^{(k)})\ . \tag{3.186}$$

Introduzindo a notação $\Delta x^{(k)} = x^{(k+1)} - x^{(k)}$, e multiplicando ambos os membros de (3.186), à esquerda, por $J_F(x^{(k)})$, obtém-se

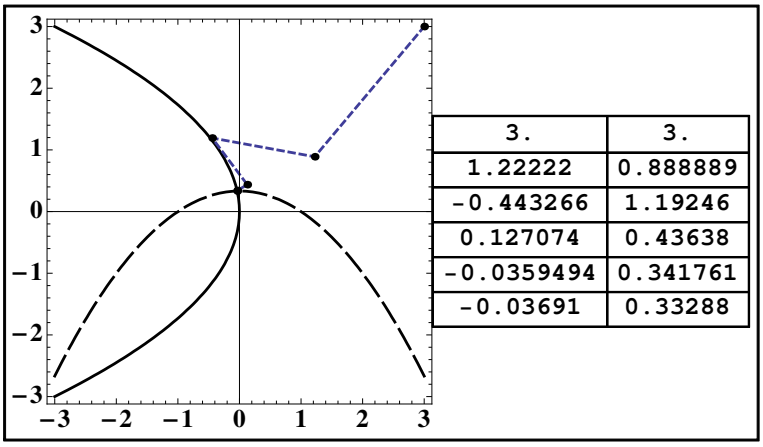
$$J_F(x^{(k)})\Delta x^{(k)} = -F(x^{(k)})\ . \tag{3.187}$$

A fórmula anterior, juntamente com

$$x^{(k+1)} = \Delta x^{(k)} + x^{(k)}, \tag{3.188}$$

define um processo iterativo, equivalente ao da fórmula (3.186), mas onde não aparece a inversa da jacobiana.

Em vez de se inverter a matriz jacobiana de F, basta-nos em cada iteração resolver o sistema linear (3.187), cuja matriz é essa jacobiana. Este sistema linear pode ser resolvido por qualquer dos métodos directos ou iterativos que estudámos nas secções anteriores. Como sabemos, a sua resolução necessita de menos operações do que a inversão da matriz correspondente.



3.	3.
1.22222	0.888889
-0.443266	1.19246
0.127074	0.43638
-0.0359494	0.341761
-0.03691	0.33288

Figura 3.6: Método de Newton para Exemplo 3.20, com $x^{(0)} = (3, 3)$.

Uma vez resolvido o sistema (3.187), a sua solução $\Delta x^{(k)}$ dá-nos a "correcção" que, somada à iterada anterior, permite obter a nova iterada $x^{(k+1)}$ (ver (3.188)).

O processo é repetido até que se verifique uma das seguintes condições (ou ambas):

$$\|\Delta x^{(k)}\| < \epsilon, \qquad \|F(x^{(k)})\| < \epsilon,$$

sendo ϵ uma margem de erro previamente fixada. Nas condições de paragem anteriores pode usar-se qualquer das normas vectoriais anteriormente estudadas.

Exemplo 3.20. *Consideremos de novo o sistema de duas equações*

$$\left\{ \begin{array}{rcrcl} 3\,x_1 & + & x_2^2 & = & 0 \\ x_1^2 & + & 3\,x_2 & = & 1\ . \end{array} \right. \tag{3.189}$$

Partindo da aproximação inicial $x^{(0)} = (0, 0)$, vamos efectuar duas iterações do método de Newton para aproximar a sua solução.

Temos

$$\begin{array}{l} F_1(x_1, x_2) = 3\,x_1 + x_2^2, \\ F_2(x_1, x_2) = x_1^2 + 3\,x_2 - 1\ . \end{array}$$

A matriz jacobiana de F é

$$J_F(x_1, x_2) = \left[\begin{array}{cc} 3 & 2x_2 \\ 2x_1 & 3 \end{array} \right]. \tag{3.190}$$

Assim, para a primeira iteração, temos

$$J_F(x^{(0)})\Delta x^{(0)} = -F(x^{(0)}), \tag{3.191}$$

onde

$$J_F(x^{(0)}) = \left[\begin{array}{cc} 3 & 0 \\ 0 & 3 \end{array} \right],$$

e

$$F(x^{(0)}) = (F_1(0,0), F_2(0,0)) = (0,-1) \ .$$

Resolvendo o sistema (3.191), obtém-se

$$\Delta x^{(0)} = (0, 1/3) \ .$$

Logo,

$$x^{(1)} = \Delta x^{(0)} + x^{(0)} = (0, 1/3) \ .$$

Passemos à segunda iteração, a qual será calculada a partir do sistema linear

$$J_F(x^{(1)})\, \Delta x^{(1)} = -F(x^{(1)}), \tag{3.192}$$

onde

$$J_F\left(x^{(1)}\right) = \begin{bmatrix} 3 & 2/3 \\ 0 & 3 \end{bmatrix},$$

e

$$F(x^{(1)}) = (F_1(0,1/3), F_2(0,1/3)) = (1/9, 0) \ .$$

Resolvendo o sistema (3.192), obtém-se

$$\Delta x^{(1)} = (-1/27, 0).$$

Finalmente, resulta a segunda iterada,

$$x^{(2)} = \Delta x^{(1)} + x^{(1)} = (-1/27, 1/3) \ . \tag{3.193}$$

Note-se que embora nos cálculos acima efectuados as duas primeiras iterações do método de Newton coincidam com as do método do ponto fixo, isto não é o que acontece em geral. Em regra, tal como acontece no caso de $n = 1$, o método de Newton, quando converge, define uma sucessão de aproximações de *convergência quadrática*, enquanto o método do ponto fixo apresenta apenas *convergência linear*. Assim, de uma maneira geral, o método de Newton, com o mesmo número de iterações, permite atingir um resultado mais preciso.

Convida-se o leitor a refazer os cálculos, começando com $x^{(0)} = (3, 3)$. Na Figura 3.6 encontram-se representados os pontos de $[-3, 3] \times [-3, 3]$ que satisfazem a equação $3\, x_1 + x_2^2 = 0$ (a negro) e a equação $x_1^2 + 3\, x_2 - 1 = 0$ (a tracejado), bem como uma tabela dando os valores aproximados das primeiras 5 iteradas do método de Newton, começando em $x^{(0)}$.

A solução z do sistema (3.189) tem por componentes

$$\begin{aligned} z_1 &= -0.036936048808669737428443360298789065613 95 \\ z_2 &= 0.332878576099467855623481498241619245764 5 \ . \end{aligned}$$

Todos os dígitos das componentes de z são significativos, e foram obtidos recorrendo ao sistema *Mathematica* [38], usando precisão arbitrária.

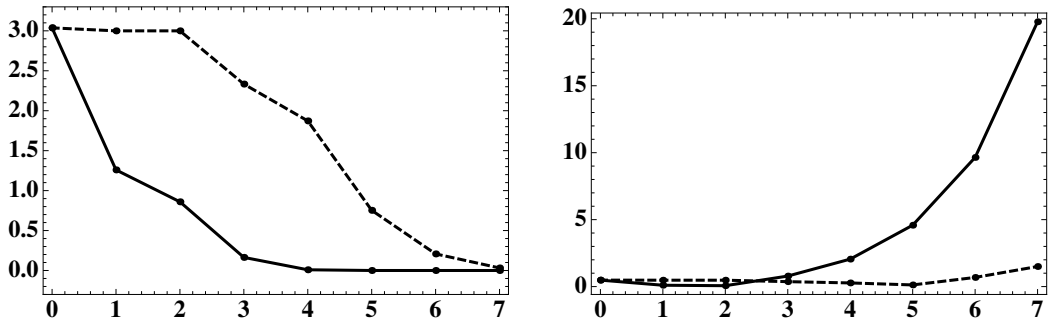


Figura 3.7: Comparação do método de Newton com um método de ponto fixo (ver Exemplo 3.20).

Efectuaram-se 7 iterações respectivamente do método de Newton e do método de ponto fixo, aplicados aos sistema (3.189).

Na Figura 3.7, do lado esquerdo, compara-se o erro calculado $||z - x^{(k)}||_\infty$, para $k = 0 : 7$, para cada um dos métodos referidos. O erro (em norma) do método de Newton (linha a cheio) decresce muito rapidamente de iterada para iterada, enquanto que para o método de ponto fixo a diminuição do erro processa-se lentamente.

A noção de número de algarismos significativos de uma aproximação (ver Definição 1.4, pág. 19), encontra neste contexto uma aplicação valiosa. Com efeito, uma vez que, dada uma aproximação $\bar{a}$ do número exacto a, tal que o respectivo erro absoluto satisfaça $0 < |a - \bar{a}| \simeq 10^{-k}$, o número de algarismos significativos de $\bar{a}$ é dado (aproximadamente) pelo valor $Sig(\bar{a})$,

$$Sig(\bar{a}) = |\log_{10}(|a - \bar{a}|)| \simeq k \ . \tag{3.194}$$

A função Sig foi aplicada, componente a componente, sobre os erros absolutos de cada aproximação $x^{(k)}$, respectivamente para cada um dos referidos métodos.

O resultado encontra-se no gráfico à direita da Figura 3.7. Note que a partir da terceira iteração o método de Newton aproximadamente *duplica* o número de algarismos significativos das componentes dos vectores de iteração deste método (linha de traço cheio), enquanto que para o método de ponto fixo, o crescimento de $Sig(\bar{x}^{(k)})$ é lento. De facto, a sétima iteração do método de Newton possui cerca de 20 algarismos significativos, enquanto a correspondente iteração do método de ponto fixo tem aproximadamente um décimo dessa precisão.

♦

3.8 Exercícios resolvidos

Os métodos iterativos de Jacobi e de Gauss-Seidel, se convergentes, produzem aproximações da solução de um sistema linear e só excepcionalmente conduzem à solução exacta do sistema. Tal acontece no caso particular de um sistema $A\,x = b$, onde a matriz dos coeficientes é *triangular superior*. O exercício a seguir ilustra este caso.

Exercício 3.3. *Considere um sistema linear $A\,x = b$, onde $A \in \mathbb{R}^{n\times n}$ é matriz (não singular) triangular superior.*

(a) Desprezando erros de arredondamento, mostre que tanto o método de Jacobi como de Gauss-Seidel produzem a solução do sistema, quando muito em n iterações.

(b) Supondo que α, β e γ são valores não nulos, aplique os métodos anteriormente referidos para determinar a solução exacta do sistema

$$\begin{bmatrix} \alpha & 1 & 1 \\ 0 & \beta & 1 \\ 0 & 0 & \gamma \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \alpha + 2 \\ \beta + 1 \\ \gamma \end{bmatrix},$$

partindo de um vector inicial qualquer $x^{(0)} = (x_{1,0}, x_{2,0}, x_{3,0})^T$.

(a) Dado que na decomposição regular da matriz A (ver pág. 134), a matriz L é a matriz nula, ou seja, $A = D + L + U = D + U$, a matriz de iteração de cada um dos métodos é da forma

$$C_J = -D^{-1}\,(L + U) = -D^{-1}\,U,$$

e

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

Assim os referidos métodos, quando aplicados ao sistema triangular dado, possuem a mesma matriz de iteração, ou seja, são o mesmo processo iterativo cuja matriz de iteração, C, é da forma

$$C = -D^{-1}\,U = \begin{bmatrix} 0 & -a_{12}/a_{11} & \cdots & -a_{1n}/a_{11} \\ 0 & 0 & \cdots & -a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

O espectro de C é constituído pelo valor próprio $\lambda = 0$ (de multiplicidade n). Logo, a respectiva equação característica é

$$(-1)^n \lambda^n = 0\ .$$

O Teorema de Cayley[23]- Hamilton[24] diz-nos que qualquer matriz quadrada é solução do respectivo polinómio característico ([27], pág. 509). Aplicando este resultado à matriz C, resulta

$$C^n = O,$$

onde O representa a matriz nula $n \times n$.

Seja $x = A^{-1}\, b$ a solução do sistema triangular dado. Partindo de um qualquer vector inicial $x^{(0)}$, sabemos que as iteradas do método satisfazem as relações de erro

$$x - x^{(k+1)} = C\, \left(x - x^{(k)}\right), \qquad k = 0, 1, 2, \ldots\ .$$

Assim,

$$\begin{aligned} x - x^{(1)} &= C\, \left(x - x^{(0)}\right) \\ x - x^{(2)} &= C\, \left(x - x^{(1)}\right) = C^2\, \left(x - x^{(0)}\right)^2, \end{aligned}$$

donde se pode concluir que, para qualquer $k \geq 1$,

$$x - x^{(k)} = C^k\, \left(x - x^{(0)}\right).$$

Por conseguinte, para $k = n$, obtém-se

$$x - x^{(n)} = C^n\, \left(x - x^{(0)}\right) = O\, \left(x - x^{(0)}\right) = 0 \Longrightarrow x = x^{(n)}.$$

A última igualdade significa que o processo iterativo produz a solução exacta x, quando muito em n iterações.

(b) As fórmulas computacionais do método podem escrever-se directamente a partir do sistema dado. Essas fórmulas definem o processo iterativo $x^{(k+1)} = -D^{-1}\, U\, x^{(k)} + D^{-1}\, b$ seguinte:

$$x^{(k+1)} = \begin{cases} x_1^{(k+1)} = \dfrac{2 + \alpha - \left(x_2^{(k)} + x_3^{(k)}\right)}{\alpha} \\ \\ x_2^{(k+1)} = \dfrac{1 + \beta - x_3^{(k)}}{\beta}, \qquad\qquad k = 0, 1, \ldots \\ \\ x_3^{(k+1)} = 1\ . \end{cases}$$

Primeira iteração:

$$x^{(1)} = \begin{cases} x_1^{(1)} = \dfrac{2 + \alpha - (x_{0,2} + x_{0,3})}{\alpha} \\ \\ x_2^{(1)} = \dfrac{1 + \beta - x_{0,3}}{\beta} \\ \\ x_3^{(1)} = 1\ . \end{cases}$$

[23] Arthur Cayley, 1821 – 1895, matemático britânico.
[24] William Rowan Hamilton, 1805 – 1865, físico, astrónomo e matemático irlandês.

Note que caso o vector inicial $x^{(0)} = (x_{1,0}, x_{2,0}, x_{3,0})$ for tal que $x_{0,2} + x_{0,3} = 2$ e $x_{0,3} = 1$, basta uma iteração para se obter a solução exacta do sistema $x = (1,1,1)^T$.

Segunda iteração:

$$x^{(2)} = \begin{cases} x_1^{(2)} = \dfrac{2 + \alpha - \dfrac{(1 + \beta - x_{0,3})}{\beta} - 1}{\alpha} = \dfrac{\alpha\,\beta - 1 + x_{0,3}}{\alpha\,\beta} \\ x_2^{(2)} = \dfrac{1 + \beta - 1}{\beta} = 1 \\ x_3^{(2)} = 1\ . \end{cases}$$

Terceira iteração:

$$x^{(3)} = \begin{cases} x_1^{(3)} = \dfrac{\alpha\,\beta - 1 + 1}{\alpha\,\beta} = 1 \\ x_2^{(3)} = 1 \\ x_3^{(3)} = 1. \end{cases}$$

Assim, a terceira iterada $x^{(3)}$ coincide com a solução $x = (1,1,1)^T$ do sistema dado. ♦

3.9 Leituras recomendadas

R. Bagnara, *A unified proof for the convergence of Jacobi and Gauss-Seidel methods*, SIAM Rev. 37, No. 1, 93-97, 1995.

Joseph F. Grcar, *Mathematicians of Gaussian Elimination*, Notices of the AMS, Vol. 58, 6, 2011.

Niall Madden, *John Todd and the Development of Modern Numerical Analysis*, Irish Math. Soc. Bulletin, 69, 11-23, 2012,
http://www.maths.tcd.ie/pub/ims/bull69/Madden.pdf.

Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.

H. Pina, *Métodos Numéricos*, Escolar Editora, 2010., Cap. 6.

David M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971, Ch. 3.

Capítulo 4

Aproximação de funções

Um dos métodos clássicos de aproximação de funções é a chamada *interpolação*, de que nos ocuparemos neste capítulo. A técnica de interpolação é muito utilizada, por exemplo, em desenho assistido por computador e na aproximação de soluções de equações diferenciais ordinárias ou às derivadas parciais.

4.1 Interpolação polinomial

Para funções reais de variável real, o objectivo da interpolação é reconstruir num certo intervalo $[a, b]$ uma função f, cujos valores são conhecidos apenas num número finito de pontos desse intervalo. Esses pontos são os chamados *nós de interpolação* e vamos representá-los genericamente por x_i $(i = 0 : n)$. Assim, os dados são constituídos por uma tabela de $n + 1$ valores de f a que chamaremos o *suporte de interpolação*,

x_0	x_1	$\ldots$	x_n
f_0	f_1	$\ldots$	f_n

onde $f_i = f(x_i)$ representa o valor de f no nó de interpolação x_i. Supomos que os nós são distintos, isto é, $x_i \neq x_j$ para $i \neq j$.

Para exemplificarmos através de uma aplicação simples, consideremos a Tabela 4.1 a seguir, que representa os valores da população de uma determinada espécie (N_i, em milhares), determinados em instantes distintos t_i.

Suponhamos que o nosso objectivo é reconstruir a função $N(t)$, descrevendo a população da espécie considerada no intervalo $[10, 16]$. Claro que, de um modo

t_i	10	12	14	16
N_i	10	15	22	18

Tabela 4.1: Valores da população de uma determinada espécie.

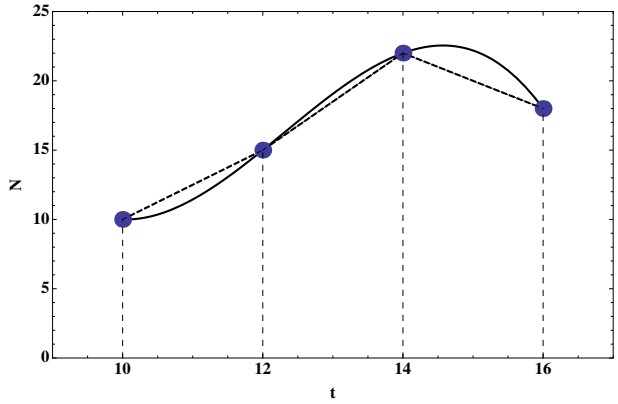


Figura 4.1: Função interpoladora N_1 (tracejado) e função interpoladora N_2 (cheio).

geral, existem muitas maneiras de "interpolar" estes dados. Na Figura 4.1 estão representadas duas das possíveis funções interpoladoras. Se usarmos a função interpoladora N_1, por exemplo, a estimativa para a população no momento $t = 15$, é $N_1(15) = 20$, enquanto que se usarmos a função N_2, o valor correspondente é $N_2(15) = 22.188$, conforme poderá verificar depois de saber construir os polinómios interpoladores N_1 e N_2.

Com efeito, iremos estudar um único tipo de interpolação, a chamada *interpolação polinomial*, o que significa que iremos considerar apenas funções interpoladoras do tipo polinomial.

No próximo parágrafo definiremos o conceito de polinómio interpolador, e demonstraremos a sua existência e unicidade.

4.1.1 Existência e unicidade do polinómio interpolador

Começamos por formular a definição de polinómio interpolador.

Definição 4.1. Fixado o número inteiro $n \geq 0$, chama-se polinómio interpolador no suporte

$$\{(x_0, f_0), \ldots, , (x_n, f_n)\},$$

ao polinómio P_n, de *grau menor ou igual* a n, que satisfaz as relações

$$P_n(x_i) = f_i, \qquad 0, 1, \ldots, n$$

A primeira questão que se põe é saber se, dado um determinado suporte, existe sempre um polinómio interpolador e se este é único.

No caso de dois nós (x_0, x_1), é simples responder a esta questão. Com efeito, segundo a Definição 4.1, o polinómio interpolador possui grau menor ou igual a um, ou seja, é uma função linear. Como o gráfico de tal função é uma recta,

é óbvio que o polinómio interpolador existe e é único – trata-se de uma função polinomial $P_1(x) = a_0 + a_1\, x$, tendo como gráfico a recta que passa pelos pontos (x_0, f_0) e (x_1, f_1).

Quando se considera um número de nós arbitrário, ou seja $n+1$ nós, o problema já não é tão simples, mas a resposta ao problema de existência e unicidade do respectivo polinómio interpolador continua a ser positiva.

Para analisarmos o caso geral, recordemos que um polinómio de grau não superior a n pode ser escrito na forma

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \tag{4.1}$$

onde os coeficientes a_i $(i = 0 : n)$ são certos números reais. Os números $a_0, \ldots, a_n$ são simplesmente as coordenadas do polinómio P_n na base[1]

$$\{1, x, x^2, \ldots x^n\},$$

do espaço linear dos polinómios de grau $\leq n$, o qual passamos a designar por $\mathcal{P}_n$.

Assim, construir o polinómio interpolador equivale a calcularmos as suas coordenadas a_i na referida base. Recorrendo de novo à definição de polinómio interpolador para o suporte $\{(x_0, f_0), \ldots, (x_n, f_n)\}$, o polinómio P_n satisfaz as igualdades

$$\begin{array}{ll} P_n(x_0) & = a_0 + a_1\, x_0 + a_2\, x_0^2 + \cdots + a_n\, x_0^n = f_0 \\ P_n(x_1) & = a_0 + a_1\, x_1 + a_2\, x_1^2 + \cdots + a_n\, x_1^n = f_1 \\ & \vdots \\ P_n(x_n) & = a_0 + a_1\, x_n + a_2\, x_n^2 + \cdots + a_n\, x_n^n = f_n\ . \end{array} \tag{4.2}$$

Observando as relações (4.2), verificamos que elas formam um sistema de $n+1$ equações lineares nas incógnitas a_0,a_1,...,a_n. Escrevendo esse sistema na forma matricial, obtém-se

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \ldots \\ f_n \end{bmatrix}. \tag{4.3}$$

A matriz deste sistema é conhecida como matriz de Vandermonde[2]. Vamos representá-la por $V(x_0, x_1, \ldots, x_n)$. Para nos certificarmos de que o sistema (4.3) tem sempre solução única, basta verificar que esta matriz é invertível, ou seja, que o seu determinante é diferente de zero.

[1] Esta base é habitualmente designada por base *canónica*.

[2] Alexandre -Théophile Vandermonde, 1735 -1796, matemático, químico e músico francês.

O caso $n = 0$ é trivial porquanto $P_0(x) = f_0$ é função interpoladora do suporte $\{x_0, f_0\}$.

Seja $n = 1$. É evidente que

$$\det(V(x_0, x_1)) = x_1 - x_0 \neq 0,$$

já que admitimos que os nós de interpolação são distintos.

Passando ao caso geral, pretendemos mostrar que é não nulo o determinante $\det(V(x_0, x_1, \ldots, x_n)) \neq 0$, para $n = 1, 2, \ldots$.

Pode provar-se que

$$\det(V(x_0, x_1, \ldots, x_n)) = \prod_{i,j=0,\, i>j}^{n} (x_i - x_j), \tag{4.4}$$

onde no produto se consideram todos os pares x_i, x_j, tais que $i > j$ (ver, por exemplo, [31], pág. 77). Conclui-se que o determinante da matriz de Vandermonde é não nulo, para qualquer n, e por conseguinte o sistema (4.3) tem sempre uma única solução (desde que os nós de interpolação sejam todos distintos).

Assim, dada uma qualquer tabela de valores de uma função f num conjunto de $n + 1$ nós distintos, existe um único polinómio interpolador.

A determinação do polinómio interpolador a partir do sistema de Vandermonde (4.3) não é todavia usada na prática, por duas ordens de razões. A primeira reside no facto de podemos obter o polinómio interpolador usando algoritmos mais económicos do ponto de vista do número de operações envolvidas. A segunda é que o sistema de Vandermonde referido pode ser extremamente mal condicionado, conforme se mostra no Exemplo a seguir.

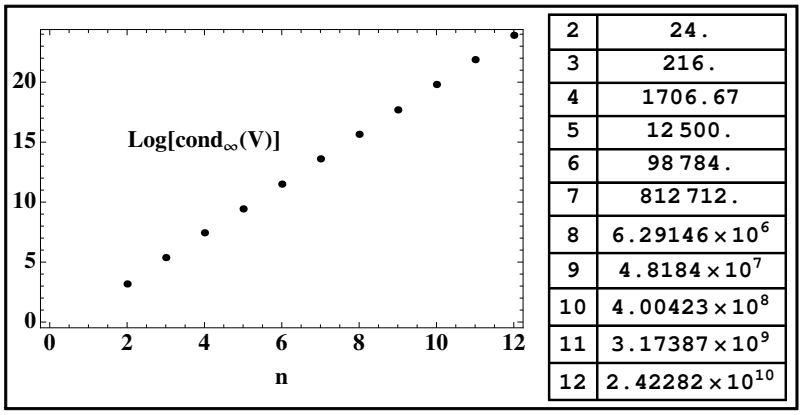
Exemplo 4.1. *Fixado $n \geq 1$, se dividirmos o intervalo $[0, 1]$ em n partes iguais, de comprimento $h = 1/n$, obtemos o suporte de interpolação*

$$x_0 = 0, \quad x_1 = 1/n, \quad x_2 = 2/n, \quad \ldots, \quad x_n = 1\ . \tag{4.5}$$

O sistema de Vandermonde (4.3) *é mal condicionado para este suporte de interpolação.*

Para n desde 2 a $n = 12$, mostra-se na Figura 4.2 a evolução do número de condição da matriz de Vandermonde correspondente, ou seja para a matriz $V = V(0, 1/n, \ldots, 1)$, na norma $||\ .||_\infty$, pág. 99.

Para evidenciarmos ser exponencial o crescimento desse número de condição, é mostrado o gráfico de $\ln(cond_\infty(V))$, em função de n. O gráfico é acompanhado de uma tabela contendo os valores de $cond_\infty(V)$, para cada valor de n considerado. Constata-se que o número de condição é muito elevado, mesmo para valores moderados de n. Assim, o sistema (4.3) associado ao suporte de interpolação em causa é extremamente mal condicionado, pelo que não deverá ser usado para calcular o polinómio interpolador de um suporte contendo os nós (4.5). ♦



2	24.
3	216.
4	1706.67
5	12 500.
6	98 784.
7	812 712.
8	6.29146×10^6
9	4.8184×10^7
10	4.00423×10^8
11	3.17387×10^9
12	2.42282×10^{10}

Figura 4.2: Variação de $cond(V)_\infty$ para n desde 2 a 12.

4.1.2 Fórmula interpoladora de Lagrange

Uma vez esclarecida a questão teórica da existência de polinómio interpolador, põe-se a questão prática de o calcular. São conhecidas fórmulas diversas para obtenção do polinómio interpolador. Iremos deduzir duas delas e compará-las do ponto de vista da sua eficiência computacional.

Nunca é demais lembrar que o polinómio interpolador de um dado suporte de interpolação é único (como se provou no parágrafo 4.1.1). Por isso, independentemente do algoritmo que usarmos para o construir, o polinómio final será sempre o mesmo.

Assim, desprezando eventuais erros de arredondamento o valor calculado do polinómio interpolador num ponto deverá ser o mesmo para qualquer fórmula interpoladora que usemos. No entanto, como sabemos, fórmulas algebricamente equivalentes podem ter comportamentos muito diversos no que toca a propagação de erros. É por conseguinte importante, neste contexto, adoptar fórmulas computacionalmente estáveis.

Interpolação de Lagrange

Uma das fórmulas mais simples para a construção do polinómio interpolador é a *fórmula interpoladora de Lagrange*. Esta fórmula baseia-se no facto de que os polinómios de grau não superior a n constituem um espaço linear de dimensão $n+1$ (o espaço linear $\mathcal{P}_n$, para a adição usual de funções e a multiplicação de uma função por um escalar). Assim, se fixarmos $n+1$ polinómios de grau não superior a n, linearmente independentes, qualquer outro polinómio de $\mathcal{P}_n$ se exprime como uma combinação linear dos polinómios fixados.

No método de Lagrange, para se construir o polinómio interpolador começamos por definir $n+1$ polinómios, que formam uma base em $\mathcal{P}_n$, designada por *base*

de Lagrange. Vamos representar esses polinómios por $L_i(x)$, $(i = 0, 1, \ldots, n)$, e designá-los como *polinómios de Lagrange*.

Os polinómios de Lagrange possuem a particularidade de serem todos *de grau exactamente n*. São construídos para um dado conjunto de nós distintos x_i, para $i = 0 : n$, de tal modo que é natural estabelecer uma correspondência entre cada nó x_i e o polinómio L_i. Esta correspondência estabelece-se do modo que é descrito a seguir.

Designamos por L_i o polinómio de grau n, tal que

$$L_i(x_i) = 1 \quad \text{e} \quad L_i(x_j) = 0, \quad \text{se} \quad j \in \{0, 1, ..., n\}, \quad \text{com} \quad j \neq i \ . \tag{4.6}$$

Como construir tal polinómio? Uma vez que ele se anula nos pontos x_0, x_1, ... , x_{i-1}, $x_{i+1}, \ldots, x_n$, é fácil concluir que tal polinómio deverá ter a forma

$$L_i(x) = A_i(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n), \qquad i = 0 : n \ . \tag{4.7}$$

onde A_i é uma certa constante real (não dependente de x). Para definir o valor desta constante, basta ter em conta a condição $L_i(x_i) = 1$. De acordo com (4.7), temos

$$L_i(x_i) = A_i(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) = 1, \qquad i = 0 : n, \tag{4.8}$$

donde

$$A_i = \frac{1}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}, \qquad i = 0 : n \ . \tag{4.9}$$

Substituindo (4.9) na expressão (4.7), obtém-se

$$\begin{aligned} L_i(x) &= \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)} \\ &= \frac{\prod_{j=0, j\neq i}^{n} (x - x_j)}{\prod_{j=0, j\neq i}^{n} (x_i - x_j)}, \quad i = 0 : n \ . \end{aligned} \tag{4.10}$$

É óbvio que cada uma das funções L_i é um polinómio de grau n. Para provar que estes polinómios formam uma base de $\mathcal{P}_n$, vamos verificar que constituem um conjunto de $n + 1$ funções linearmente independentes.

Considere-se uma combinação linear da forma

$$\sum_{j=0}^{n} c_j L_j(x), \tag{4.11}$$

onde c_j são coeficientes reais arbitrários. Devemos provar que

$$\sum_{j=0}^{n} c_j L_j(x) \equiv 0, \quad \forall x \in \mathbb{R}, \tag{4.12}$$

se e só se $c_0 = c_1 = \cdots = c_n = 0$. Temos,

$$\sum_{j=0}^{n} c_j \, L_j(x_i) = 0 \Longleftrightarrow c_i \, L_i(x_i) = 0 \Longleftrightarrow c_i = 0 \ .$$

Conclui-se que $c_i = 0$, para $i = 0, 1, .., n$, isto é, a identidade (4.12) só se verifica se todos os coeficientes c_i se anularem simultaneamente. Logo, os $n+1$ polinómios de Lagrange são linearmente independentes, pelo que formam uma base de $\mathcal{P}_n$. A esta base chamamos a *base de Lagrange* associada aos nós $x_0, x_1, \ldots, x_n$.

Por conseguinte, dada uma tabela de valores de uma certa função f nos pontos x_i, o polinómio interpolador de f nesses pontos pode ser representado (de forma única) como

$$P_n(x) = \sum_{j=0}^{n} d_j L_j(x) \ . \tag{4.13}$$

Resta-nos determinar as coordenadas d_j do polinómio interpolador na base de Lagrange, o que é bastante fácil tendo em conta a definição dos polinómios de Lagrange.

Com efeito, para que o polinómio P_n dado em (4.13) seja o polinómio interpolador de f, basta escolher $d_j = f(x_j)$, para $j = 0, ..., n$. Isto é, considerar a seguinte combinação linear dos elementos que constituem a base de Lagrange,

$$P_n(x) = \sum_{j=0}^{n} f(x_j) L_j(x) \ . \tag{4.14}$$

Para provarmos a validade da fórmula (4.14), basta recordar a definição dos polinómios de Lagrange. De facto, calculando P_n em x_i, e usando (4.6), a fórmula (4.14) reduz-se a

$$P_n(x_i) = \sum_{j=0}^{n} f(x_j) L_j(x_i) = f(x_i) L_i(x_i) = f(x_i), \qquad i = 0 : n \ . \tag{4.15}$$

A igualdade (4.15) é satisfeita em todos os nós x_i e portanto comprova-se que o polinómio P_n, definido por (4.14), é o *polinómio interpolador de f* nestes nós, uma vez que o polinómio interpolador é único.

A fórmula (4.14) é conhecida como *fórmula interpoladora de Lagrange*, sendo os polinómios da base de Lagrange definidos por (4.10).

4.1.3 Escolha dos nós de interpolação

Por vezes, ao resolver um problema mediante aplicação de um determinado método dispomos de informação redundante. Por exemplo, se quisermos aproximar uma função por um polinómio de grau 2 e conhecermos os seus valores em quatro pontos é óbvio que teremos de descartar um dos pontos.

São possíveis vários critérios para a selecção dos nós de interpolação. Em primeiro lugar, se a função considerada apresentar uma ou mais descontinuidades, faz sentido aproximá-la por troços. Isto é, se por exemplo ela for descontínua em $x = 0$, a interpolação deve ser feita separadamente para valores de x positivos e negativos. Não faria sentido aproximar uma tal função usando dois nós de sinais opostos.

Se não for este o caso, isto é, se a função a interpolar for contínua em todo o domínio considerado, então o critério mais comum para a escolha dos nós de interpolação é a *proximidade*. Isto é, se quisermos aproximar a função num certo ponto x, devem escolher-se primeiro os dois pontos mais próximos de x, sendo os pontos seguintes escolhidos pelo mesmo critério. Embora o erro de interpolação, como veremos mais adiante, dependa de vários factores, na ausência de outras informações sobre a função, esta é a melhor escolha possível para o minimizar.

No parágrafo 4.1.9, pág. 207, referir-nos-emos a outros possíveis critérios de escolha dos pontos, relacionados com a minimização do erro de interpolação.

Exemplo 4.2. *Consideremos a função, dada pela tabela numérica 4.1, pág. 183. O nosso objectivo é obter valores aproximados de* $N(15)$ *(valor da população no instante* $t = 15$*), por interpolação polinomial, aplicando a fórmula interpoladora de Lagrange.*

(a) Utilizando interpolação linear.

(b) Utilizando interpolação quadrática (ou parabólica).

(c) Usando todos os pontos da tabela.

(a) Para se aplicar interpolação linear (isto é, utilizando um polinómio de grau não superior a 1), devemos considerar os valores de N em dois pontos. De acordo com o que se disse anteriormente, os pontos deverão ser os nós mais próximos de $x = 15$, ou seja, $x_0 = 14$ e $x_1 = 16$. Note-se que a ordem dos pontos escolhidos é arbitrária, não influindo no resultado da interpolação.

Seja P_1 o polinómio que interpola a função N em x_0 e x_1. Para o calcularmos, começamos por construir a respectiva base de Lagrange. De acordo com a fórmula (4.10), pág. 188, temos

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x - 16}{-2}, \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - 14}{2} \ .$$

Aplicando a fórmula interpoladora de Lagrange (4.14), resulta

$$P_1(x) = f(x_0)\, L_0(x) + f(x_1)\, L_1(x) = 22\ \frac{x-16}{-2} + 18\ \frac{x-14}{2}\ .$$

A aproximação desejada é $P_1(15) = 11 + 9 = 20$.

(b) No caso de interpolação quadrática são necessários 3 nós de interpolação. Usando de novo o critério de proximidade, o terceiro ponto a considerar é $x_2 = 12$. Os polinómios de Lagrange correspondentes são,

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-16)(x-12)}{(-2)(2)}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-14)(x-12)}{(2)(4)}$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-14)(x-16)}{(-2)(-4)}\ .$$

Aplicando a fórmula interpoladora de Lagrange (4.14), temos

$$\begin{aligned} P_2(x) \ &= f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2(x) \\ &= 22\ \frac{(x-16)(x-12)}{-4} + 18\ \frac{(x-14)(x-12)}{8} + 15\ \frac{(x-14)(x-16)}{8}\ . \end{aligned}$$

Donde, $P_2(15) = 22 \times 3/4 + 18 \times 3/8 + 15 \times (-1/8) = 21.375$.

(c) Se usarmos todos os pontos da tabela, estaremos a fazer interpolação cúbica (de grau 3). Uma vez que a ordem dos pontos é irrelevante para o resultado, designemos por x_3 o ponto $x_3 = 10$, mantendo as designações dos restantes pontos de interpolação.

Os polinómios de Lagrange correspondentes são,

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-16)(x-12)(x-10)}{(-2)(2)(4)}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x-14)(x-12)(x-10)}{(2)(4)(6)}$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x-14)(x-16)(x-10)}{(-2)(-4)2}$$

$$L_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x-14)(x-16)(x-12)}{(-2)(-6)(-4)}\ .$$

Aplicando a fórmula interpoladora de Lagrange, temos

$$\begin{aligned} P_3(x) &= f(x_0)\,L_0(x) + f(x_1)\,L_1(x) + f(x_2)L_2(x) + f(x_3)L_3(x) \\ &= 22\,\frac{(x-16)(x-12)(x-10)}{-16} + 18\,\frac{(x-14)(x-12)(x-10)}{48} + \\ &+15\,\frac{(x-14)(x-16)(x-10)}{16} + 10\,\frac{(x-14)(x-16)(x-12)}{-48}\ . \end{aligned}$$

A aproximação do valor da função em 15, por interpolação cúbica, é $P_3(15) = 22.1875$. ♦

4.1.4 Fórmula interpoladora de Newton

No Exemplo 4.2, a fórmula interpoladora de Lagrange foi utilizada para calcular uma sucessão de aproximações do valor da função em causa num ponto do seu argumento. Evidencia-se desde logo uma desvantagem computacional da fórmula de Lagrange – de cada vez que se determina uma nova aproximação, aumentando o grau do polinómio interpolador, é necessário recalculá-la sem aproveitar os cálculos efectuados anteriormente.

Vamos estudar uma fórmula de cálculo alternativa, em que o polinómio interpolador de um certo grau é obtido como uma *correcção* do polinómio do grau anterior. Este método (conhecido como *fórmula interpoladora de Newton*) é mais eficiente, diminuindo substancialmente o número total de operações aritméticas necessárias e consequentemente providenciando grealmente uma fórmula numericamente mais estável do que a da interpoladora de Lagrange.

Para estudar a fórmula interpoladora de Newton, comecemos por formular o seguinte problema. Seja P_n o polinómio de grau menor ou a igual a n que interpola uma certa função f nos nós x_0,x_1, ..., x_n. Se ao suporte considerado acrescentarmos mais um nó, seja x_{n+1}, o resultado da interpolação passará a ser o polinómio P_{n+1}, que interpola f também neste ponto.

Vamos construir o polinómio P_{n+1} a partir de P_n. Comecemos por escrever

$$P_{n+1}(x) = P_n(x) + C_{n+1}(x)\ . \tag{4.16}$$

Assumindo que P_{n+1} é diferente de P_n, sucede que C_{n+1} é geralmente um polinómio de grau $n+1$ (o mesmo grau de P_{n+1}). Facilmente se verifica que as raízes deste último polinómio coincidem com os nós de interpolação iniciais x_0,x_1,...,x_n. Com efeito, da igualdade (4.16) resulta imediatamente

$$C_{n+1}(x_i) = P_{n+1}(x_i) - P_n(x_i) = f(x_i) - f(x_i) = 0, \qquad i = 0, 1, \ldots n\ .$$

Por conseguinte, C_{n+1} pode ser escrito na forma

$$C_{n+1}(x) = A_{n+1}\,(x - x_0)\,(x - x_1)\cdots(x - x_n),$$

onde A_{n+1} não depende de x. Podemos então rescrever a fórmula (4.16) como

$$P_{n+1}(x) = P_n(x) + A_{n+1}(x - x_0)(x - x_1)\dots(x - x_n) \ . \tag{4.17}$$

O problema de calcular P_{n+1} ficou reduzido a determinar A_{n+1}, uma constante que depende dos valores de f em x_0, x_1,...,x_n. Note-se que, no caso de $P_{n+1}(x)$ coincidir com $P_n(x)$ (o que acontece se tivermos $P_n(x_{n+1}) = f(x_{n+1})$), resulta que $A_{n+1} = 0$. Se excluirmos este caso, P_{n+1} é um polinómio de grau $n + 1$, que se pode escrever na forma

$$P_{n+1}(x) = A_{n+1}x^{n+1} + \dots ,$$

ou seja, A_{n+1} é o coeficiente do termo em x^{n+1} (termo principal) de P_{n+1}.

Diferenças divididas

As considerações anteriores justificam a introdução da seguinte definição.

Definição 4.2. Chama-se *diferença dividida*, de ordem k, da função f nos nós $x_0, x_1, \dots, x_k$, ao coeficiente A_k do termo em x^k do polinómio P_k que interpola f nos nós considerados. Designa-se A_k por $f[x_0, x_1, \dots, x_k]$.

Para calcularmos diferenças divididas usa-se um processo recursivo que passamos a descrever.

Comecemos por considerar as diferenças divididas de primeira ordem, isto é, com dois nós.

Seja P_0 o polinómio que interpola f em x_0, $P_0(x) \equiv f(x_0)$. Sendo x_1 um novo ponto de interpolação, de acordo com a fórmula (4.17), o polinómio P_1, que interpola f em x_0 e x_1, é dado por

$$P_1(x) = P_0(x) + A_1(x - x_0) = f(x_0) + A_1(x - x_0) \ . \tag{4.18}$$

O valor de $A_1 = f[x_0, x_1]$ (diferença dividida de f em x_0 e x_1) deduz-se facilmente a partir da condição $P_1(x_1) = f(x_1)$. De acordo com (4.18), obtém-se

$$P_1(x_1) = f(x_0) + A_1(x_1 - x_0) = f(x_1) \ .$$

Assim,

$$A_1 = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \ . \tag{4.19}$$

No caso de $f(x_0) = f(x_1)$, temos $A_1 = 0$. Este é o único caso em que o polinómio P_1 coincide com P_0, ou seja, o respectivo polinómio interpolador com dois nós possui grau 0.

Generalizando a fórmula (4.19) para quaisquer dois nós de interpolação x_i e x_j, podemos escrever a fórmula da diferença dividida de primeira ordem,

$$f[x_i, x_j] = \frac{f(x_j) - f(x_i)}{x_j - x_i} \ .$$

A fórmula anterior justifica a designação *diferença dividida* dada ao símbolo $f[x_i, x_j]$.

A diferença dividida de primeira ordem tem um significado geométrico simples: é o declive da recta que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$ (recta cujos pontos coincidem com o gráfico do polinómio P_1).

Para se construir as diferenças divididas de ordem superior (correspondentes aos polinómios interpoladores de grau maior que um) torna-se necessário deduzir uma fórmula de recorrência.

Suponhamos que é conhecido o polinómio P_n, que interpola f em $x_0, x_1, ..., x_n$. Isto significa que é conhecida a diferença dividida $f[x_0, x_1, ..., x_n]$, já que ela é o coeficiente do termo principal de P_n, ou seja, de x^n.

Para se obter o polinómio P_{n+1} precisamos de introduzir mais um nó de interpolação, seja x_{n+1}. Em geral, temos $P_{n+1}(x_{n+1}) = f(x_{n+1})$, mas $P_n(x_{n+1}) \neq f(x_{n+1})$, já que P_{n+1} interpola f neste ponto (ao contrário de P_n).[3]

Vamos definir um polinómio auxiliar Q_n do seguinte modo: Q_n interpola f em $x_1, x_2, \ldots, x_{n+1}$. Logo, Q_n é um polinómio de grau não superior a n, tal como P_n, mas cujo termo principal tem o coeficiente $f[x_1, x_2, ..., x_{n+1}]$.

Mostre-se que P_{n+1} pode ser obtido a partir de P_n e Q_n, através da fórmula

$$P_{n+1}(x) = \frac{P_n(x)(x_{n+1} - x) + Q_n(x)(x - x_0)}{x_{n+1} - x_0} \ . \tag{4.20}$$

Para tanto, basta provar que $P_{n+1}(x_i) = f(x_i)$, para $i = 0, 1, ..., n+1$. Se $i = 0$, temos

$$P_{n+1}(x_0) = \frac{f(x_0)(x_{n+1} - x_0)}{x_{n+1} - x_0} = f(x_0) \ .$$

Por outro lado, se $i \in \{1, 2, ..., n\}$, verifica-se

$$P_{n+1}(x_i) = \frac{f(x_i)(x_{n+1} - x_i) + f(x_i)(x_i - x_0)}{x_{n+1} - x_0} = \frac{f(x_i)(x_{n+1} - x_0)}{x_{n+1} - x_0} = f(x_i) \ .$$

Finalmente, para $i = n + 1$, obtém-se

$$P_{n+1}(x_{n+1}) = \frac{f(x_{n+1})(x_{n+1} - x_0)}{x_{n+1} - x_0} = f(x_{n+1}) \ .$$

[3]Pode dar-se o caso de $P_n(x_{n+1}) = f(x_{n+1})$. Nesse caso, P_{n+1} coincide com P_n e a diferença dividida $f[x_0, x_1, ..., x_{n+1}]$ é nula.

Por conseguinte, acabamos de provar que P_{n+1}, definido pela fórmula (4.20), é o polinómio que interpola f nos pontos $x_0, x_1, ..., x_{n+1}$.

Por definição, a diferença dividida $f[x_0, x_1, ..., x_{n+1}]$ é o coeficiente do termo principal deste polinómio. Assim, ela pode ser calculada através da fórmula

$$f[x_0, x_1, ..., x_{n+1}] = \frac{f[x_1, x_2, ..., x_{n+1}] - f[x_0, x_1, ..., x_n]}{x_{n+1} - x_0}, \tag{4.21}$$

onde, como já sabemos, $f[x_0, x_1, ..., x_n]$ é o coeficiente do termo principal de P_n e $f[x_1, ..., x_{n+1}]$ é o coeficiente do termo principal de Q_n .

A fórmula (4.21) permite-nos calcular uma diferença dividida de ordem $n+1$ a partir de duas diferenças divididas de ordem n. Aplicando sucessivamente esta fórmula de recorrência, podemos calcular diferenças divididas de qualquer ordem (desde que, evidentemente, se disponha de valores suficientes da função f).

Recapitulando, para construir o polinómio interpolador P_n pela fórmula de Newton, num certo suporte de interpolação, devemos proceder do seguinte modo:

(i) Calcular as diferenças divididas de f nos pontos considerados, até à ordem n, com base na fórmula (4.21);
(ii) Determinar $P_0(x) \equiv f(x_0)$;
(iii) Obter os polinómios P_1, P_2,..., P_n, através da aplicação sucessiva da fórmula (4.17), onde A_{n+1} representa uma diferença dividida da ordem correspondente.

Este processo pode ser facilmente programado. Quando os cálculos são efectuados manualmente é costume representar as diferenças divididas numa tabela (ver Exemplo 4.3 adiante).

Base de Newton

Vimos que o polinómio interpolador anteriormente deduzido tem a forma

$$P_n(x) = c_0 + c_1\,(x-x_0) + c_2\,(x-x_0)(x-x_1) + \ldots + c_n\,(x-x_0)(x-x_1)\cdots(x-x_{n-1})\ .$$

Os coeficientes c_i, para $i = 0 : n$, são diferenças divididas construídas a partir dos nós $x_0, x_1, \ldots, x_n$. A expressão anterior de P_n significa que o polinómio interpolador possui as coordenadas c_i, na base

$$\mathcal{N} = \{1, x - x_0, (x - x_0)(x - x_1), \ldots, (x - x_0)(x - x_1)\cdots(x - x_{n-1})\}\ .$$

Esta base recebe a designação de *base de Newton*. Ela voltará a ser útil quando estudarmos algumas regras de quadratura (Capítulo 5, pág. 241).

Em resumo, o polinómio interpolador de Newton tem a forma,

$$\begin{aligned} P_n(x) &= f[x_0] + f[x_0, x_1]\,(x - x_0) + \ldots + f[x_0, x_1, \ldots, x_n]\,(x - x_0)\cdots(x - x_{n-1}) \\ &= f[x_0] + \textstyle\sum_{i=1}^{n} f[x_0, \ldots, x_i]\,(x - x_0)\ldots(x - x_{i-1})\ . \end{aligned} \tag{4.22}$$

Apresentamos a seguir alguns exemplos de aplicação da teoria exposta.

Exemplo 4.3. *Pretende-se construir a tabela de diferenças divididas correspondente à tabela 4.1, pág. 183, ordenando os nós de interpolação segundo a sua proximidade ao ponto $x = 15$ (à semelhança do Exemplo 4.2, pág. 190).*

Uma tabela de diferenças divididas pode ser estabelecida de modo "triangular". A sua construção começa pelo suporte de interpolação, ou seja, considerando duas colunas que contêm os dados do problema: uma com os valores de x_i, e outra, com os de $f(x_i) = f_i$. Na coluna seguinte, são calculadas as diferenças divididas de primeira ordem. No caso concreto da referida tabela temos 4 nós de interpolação, logo podemos calcular três destas diferenças,

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{18 - 22}{16 - 14} = -2$$

$$f[x_1, x_2] = \frac{f_2 - f_1}{x_2 - x_1} = \frac{15 - 18}{12 - 16} = \frac{3}{4}$$

$$f[x_2, x_3] = \frac{f_3 - f_2}{x_3 - x_2} = \frac{10 - 15}{10 - 12} = \frac{5}{2}\ .$$

Segue-se a coluna com as diferenças de segunda ordem,

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = -\frac{11}{8}$$

$$f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1} = -\frac{7}{24}\ .$$

Finalmente, o vértice do "triângulo" é constituído pela diferença dividida de terceira ordem,

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0} = -\frac{13}{48}\ . \tag{4.23}$$

O aspecto gráfico da tabela é o seguinte:

x_i	f_i	ordem 1	ordem 2	ordem 3
14	22			
		-2		
16	18		$-\frac{11}{8}$	
		$\frac{3}{4}$		$-\frac{13}{48}$
12	15		$-\frac{7}{24}$	
		$\frac{5}{2}$		
10	10			

A localização de cada entrada da tabela sugere-nos a maneira como se calculam as entradas sucessivas da tabela. O numerador da fracção (4.21) é a diferença entre as duas entradas adjacentes da coluna anterior; o denominador dessa fracção é a diferença entre os extremos da base do triângulo cujo vértice se encontra na entrada a calcular. ♦

Exemplo 4.4. *Retomando o Exemplo 4.2, pág. 190, pretende-se obter aproximações do valor da população, $N(15)$, usando interpolação linear, quadrática e cúbica, recorrendo à fórmula interpoladora de Newton.*

A tabela de diferenças divididas para este problema já foi calculada no exemplo anterior. Para obtermos as aproximações pedidas, basta utilizar a fórmula (4.17), pág. 193.

Dado que $P_0(x) \equiv f(x_0) = 22$, aplicando a fórmula (4.17), com $n = 0$ obtém-se o polinómio interpolador de primeiro grau,

$$P_1(x) = P_0(x) + f[x_0, x_1](x - x_0) = 22 - 2(x - 14) \ .$$

Utilizando o polinómio P_1, obtém-se a aproximação por interpolação linear,

$$N(15) \approx P_1(15) = 22 - 2(15 - 14) = 20 \ .$$

Aplicando agora a fórmula (4.17) com $n = 1$, obtém-se o polinómio interpolador de segundo grau,

$$\begin{aligned} P_2(x) &= P_1(x) + f[x_0, x_1, x_2](x - x_0)(x - x_1) = \\ &= 22 - 2(x - 14) - 11/8(x - 14)(x - 16) \ . \end{aligned}$$

Usando este polinómio, obtém-se a aproximação por interpolação quadrática,

$$N(15) \approx P_2(15) = 20 - 11/8(15 - 14)(15 - 16) = 171/8 = 21.375 \ .$$

ano	1991	1992	1993	1994	2004	2010
P	6.5	220	320	415	792.5	996.85
S	200	222.5	237	246.5	374	475

Tabela 4.2: P é o valor médio das propinas das licenciaturas (em euros) e S o salário mínimo nacional (em euros).

Finalmente, para obter o polinómio interpolador de grau 3, aplica-se a mesma fórmula com $n = 2$:

$$\begin{aligned} P_3(x) &= P_2(x) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) = \\ &= 22 - 2(x - 14) - 11/8(x - 14)(x - 16) - 13/48(x - x_0)(x - x_1)(x - x_2) \ . \end{aligned}$$

Obtém-se assim a aproximação por interpolação cúbica,

$$N(15) \approx P_3(15) = 21.375 - 13/48(15 - 14)(15 - 16)(15 - 12) = 22.1875 \ .$$

♦

Exemplo 4.5. *Na Tabela 4.2 é dada a evolução do valor médio das propinas das licenciaturas em Portugal, no período entre 1991 e 2010, bem como da evolução do salário mínimo nacional no mesmo período.*[4]

Vamos ilustrar a aplicação da interpolação polinomial respondendo às questões a seguir formuladas.

(a) Será que a evolução do valor das propinas entre 1991 *e* 1994 *permitia prever o valor que elas iriam atingir em 2004 e em 2010?*

(b) Utilizando interpolação cúbica, que previsão se pode fazer para o valor das propinas, extrapolando para 2020*?*

(c) A razão entre o valor das propinas e o salário mínimo, seja $\Pi = P/S$*, é um importante índice sobre o grau de acesso ao ensino superior. Pretende-se analisar a evolução de* Π *e, por interpolação quadrática, obter uma previsão do valor que esse índice atingirá em* 2020*.*

(d) Com base na evolução do salário mínimo nacional obtenha uma previsão do salário mínimo nacional em 2020 *(efectue extrapolação quadrática).*

(e) Com base nas respostas às duas últimas alíneas, obtenha uma nova previsão do valor das propinas em 2020*, e compare com a previsão obtida na alínea (b).*

(f) Admitindo que o valor das propinas é uma função do salário mínimo nacional, obtenha estimativas do valor das propinas quando o salário mínimo nacional atingir 550 *euros, usando extrapolação quadrática e cúbica.*

[4]Para simplificar, quando nos referirmos às propinas no ano lectivo $N/(N+1)$, consideramos que as mesmas correspondem ao ano N. Os dados referentes à evolução do salário mínimo nacional foram obtidos em http://www.dgert.mtss.gov.pt/, Direcção-Geral do Emprego e das Relações de Trabalho.

(a) Responderemos a esta pergunta por etapas.

(i) Utilizando interpolação quadrática e aplicando a fórmula de Newton, vamos obter uma estimativa do valor das propinas em 2004, ou seja $P(2004)$ (baseado apenas nos valores das propinas em anos anteriores).

Para realizarmos interpolação parabólica, usamos os valores da função P nos três anos anteriores a 2004; como devemos escolher os três anos mais próximos de 2004, temos $1992, 1993$ e 1994. Calculemos as diferenças divididas

$$\begin{array}{ll} P[1992,1993] & = (P(1993)-P(1992))/(1993-1992) = 100 \\ P[1993,1994] & = (P(1994)-P(1993))/(1994-1993) = 95 \\ P[1992,1993,1994] & = (P[1993,1994]-P[1992,1993])/(1994-1992) \\ & = -2.5 \ . \end{array}$$

Aplicando a fórmula interpoladora de Newton (4.22), pág. 196, tem-se

$$\begin{array}{rl} P_2(t) = & P(1992)+P[1992,1993]\,(t-1992)+ \\ & +P[1992,1993,1994]\,(t-1992)(t-1993) \ . \end{array}$$

Finalmente, substituindo t pelo ano em causa, 2004, obtém-se

$$P_2(2004) = 1090 \ .$$

(ii) Aplicando de novo a fórmula de Newton, vamos obter uma estimativa de $P(2010)$, também por interpolação quadrática.

Devemos basear-nos nos valores da função P nos três anos mais recentes, anteriores a 2010, ou seja, $1993, 1994$ e 2004. Calculemos as diferenças divididas

$$\begin{array}{ll} P[1994,2004] & = (P(2004)-P(1994))/(2004-1994) = 37.75 \\ P[1993,1994,2004] & = (P[1994,2004]-P[1993,1994])/(2004-1993) \\ & = -5.20364 \ . \end{array}$$

Aplicando a fórmula interpoladora de Newton, tem-se:

$$\begin{array}{rl} Q_2(t) = & P(1993)+P[1993,1994]\,(t-1993)+ \\ & +P[1993,1994,2004]\,(t-1993)(t-1994) \ . \end{array}$$

Finalmente, substituindo t pelo ano em causa, ou seja 2010, resulta

$$Q_2(2010) = 519.364 \ .$$

(ii) Comparemos as estimativas anteriormente calculadas com os valores reais dados na tabela.

A primeira estimativa, comparada com o valor de $P(2004) = 792.5$, possui um erro por excesso de 297.5, visto que $P(2004) - P_2(2004) = -297.5$. A segunda

estimativa tem um erro por defeito de aproximadamente 477.5, pois $P(2010) - Q_2(2010) \simeq 477.486$. Erros tão significativos não são de estranhar, já que neste caso estamos a fazer uma *extrapolação*, isto é, estamos a basear-nos em valores da função P em certos intervalos para obter estimativas do valor dessa função em pontos exteriores a esses intervalos.

(b) Para realizar interpolação cúbica, devemos basear-nos nos valores da função P em 1993, 1994 , 2004 e 2010. Vamos aproveitar as diferenças divididas que já calculamos não envolvendo o ano 2010. Além disso, precisamos de calcular mais três diferenças,

$$\begin{aligned} P[2004, 2010] &= (P(2010) - P(2004))/(2010 - 2004) \simeq 34.0583 \\ P[1994, 2004, 2010] &= (P[2004, 2010] - P[1994, 2004])/(2010 - 1994) \\ &\simeq -0.230729. \\ P[1993, 1994, 2004, 2010] &= \frac{P[1994, 2004, 2010] - P[1993, 1994, 2004]}{2010 - 1993} \\ &\simeq 0.292577. \end{aligned}$$

Aplicando a fórmula interpoladora de Newton, tem-se

$$P_3(t) = Q_2(t) + P[1993, 1994, 2004, 2010](t - 1993)(t - 1994)(t - 2004) \ .$$

Finalmente, substituindo t pelo ano em causa, 2020, obtém-se

$$P_3(2020) \simeq 2517.6.$$

(c) Passemos a usar a fórmula interpoladora de Lagrange (4.11), pág. 188.

Comecemos por calcular os quocientes P/S nos anos considerados:

$$\begin{aligned} P(1991)/S(1991) &= 0.0325, & P(1992)/S(1992) &= 0.988764 \\ P(1993)/S(1993) &= 1.35021, & P(1994)/S(1994) &= 1.68357 \\ P(2004)/S(2004) &= 2.11898, & P(2010)/S(2010) &= 2.09863 \ . \end{aligned}$$

Ou seja, durante o período em causa o valor das propinas passou de cerca de 3 por cento, para mais do dobro do valor do salário mínimo.

Para se fazer a interpolação pedida, tenhamos em conta os três últimos valores da tabela, correspondentes a $t_0 = 1994, t_1 = 2004$ e $t_2 = 2010$.

Os polinómios de Lagrange para estes pontos são

$$l_0(t) = \frac{(t - t_1)(t - t_2)}{(t_0 - t_1)(t_0 - t_2))} = \frac{(t - 2004)(t - 2010)}{(1994 - 2004)(1994 - 2010)}$$

$$l_1(t) = \frac{(t - t_0)(t - t_2)}{(t_1 - t_0)(t_1 - t_2))} = \frac{(t - 1994)(t - 2010)}{(2004 - 1994)(2004 - 2010)}$$

$$l_2(t) = \frac{(t - t_0)(t - t_1)}{(t_2 - t_0)(t_2 - t_1)} = \frac{(t - 1994)(t - 2004)}{(2010 - 1994)(2010 - 2004)} \ .$$

Designando por $\Pi_2(t)$ o polinómio interpolador do suporte em causa, obtém-se,

$$\begin{aligned}\Pi_2(t) &= P(t_0)/S(t_0)\, l_0(t) + P(t_1)/S(t_1)\, l_1(t) + P(t_2)/S(t_2)\, l_2(t) \\ &= 1.68357\, l_0(t) + 2.11898\, l_1(t) + 2.09863\, l_2(t)\ .\end{aligned}$$

Para responder à questão (c), basta calcular

$$\Pi_2(2020) \simeq 1.595\ .$$

(d) Aproveitando os polinómios de Lagrange anteriormente calculados, podemos escrever a seguinte fórmula para o polinómio quadrático S_2, que interpola a 'função de salários' $S(t)$, em 1994, 2004 e 2010,

$$S_2(t) = S(1994)\, l_0(t) + S(2004)\, l_1(t) + S(2010)\, l_2(t)\ .$$

Substituindo t por 2020, obtém-se a previsão pedida:

$$S_2(2020) \simeq 684.17\ .$$

(e) Neste caso, baseamo-nos na previsão do valor do salário $S_2(2020)$ e na relação propinas/salário para o mesmo ano $\Pi_2(2020) \simeq 1.595$. Obtém-se

$$P(2020) = S_2(2020) \times \Pi_2(2020) \simeq 1091.5\ .$$

(f) Devemos extrapolar os dados $P(237) = 320$, $P(246.5) = 415$, $P(374) = 996.85$ e $P(475) = 976.85$. No caso da interpolação quadrática, utilizam-se os 3 últimos valores de P (aqueles cuja abcissa é mais próxima de 550). Representando por P_2 o polinómio interpolador correspondente (que se pode obter por qualquer uma das fórmulas já utilizadas) obtém-se

$$P_2(550) = 1\,094.44\ .$$

No caso da interpolação cúbica, utilizam-se todos os pontos referidos, resultando

$$P_3(550) = 1\,890.26\ .$$

A disparidade dos resultados anteriores não é de estranhar tendo em conta que se efectuaram extrapolações.

4.1.5 Diferenças divididas como funções simétricas dos argumentos

Fixado $n \geq 1$ a diferença dividida $f[x_0, x_1, \cdots, x_n]$ goza de uma propriedade de invariância, no sentido de manter o mesmo valor para qualquer permutação que

efectuarmos do suporte $x_0, \ldots, x_n$. Por essa razão se diz que tal diferença dividida é uma *função simétrica dos seus argumentos*.
Por exemplo, para $n = 2$, o polinómio interpolador na forma de Lagrange escreve-se

$$P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \\ + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \ .$$

Por conseguinte, o coeficiente de x^2 de $P_2(x)$ é

$$\frac{f(x_0)}{(x_0-x_1)(x_0-x_2)} + \frac{f(x_1)}{(x_1-x_0)(x_1-x_2)} + \frac{f(x_2)}{(x_2-x_0)(x_2-x_1)} \ .$$

Dado que o polinómio interpolador é único e, como sabemos, o coeficiente de x^2 do polinómio P_2 quando expresso na base de Newton é $f[x_0, x_1, x_2]$, esta diferença dividida pode assim escrever-se como a soma

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)} + \frac{f(x_1)}{(x_1-x_0)(x_1-x_2)} + \frac{f(x_2)}{(x_2-x_0)(x_2-x_1)} \ . \tag{4.24}$$

Se no lugar de $f[x_0, x_1, x_2]$ considerarmos por exemplo $f[x_1, x_0, x_2]$ ou $f[x_2, x_0, x_1]$ (há $3! = 6$ maneiras de ordenarmos os nós) a expressão da soma no membro direito de (4.24) mantém-se, ou seja, podemos afirmar que $f[x_0, x_1, x_2]$ é invariante para qualquer permutação dos nós (o que é o mesmo que dizer que $f[x_0, x_1, x_2]$ é uma função simétrica dos seus argumentos). De igual modo se pode concluir que para $n \geq 1$,

$$f[x_0, x_1, \cdots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{j\neq i}^{n}(x_i - x_j)},$$

é função simétrica dos seus $n+1$ argumentos. Note-se que no produto $\prod_{j\neq i}^{n}(x_i - x_j)$ da expressão anterior, o índice i está fixado e j toma os valores de 0 a n, excluindo o valor $j = i$.

4.1.6 Erro de interpolação

Neste parágrafo vamos discutir o erro de interpolação, ou seja,

$$e_n(x) = f(x) - P_n(x),$$

onde P_n é o polinómio que interpola uma dada função f em $n+1$ nós $x_0, x_1, \ldots, x_n$.

O estudo do erro de interpolação permite-nos nomeadamente decidir qual o grau do polinómio interpolador que melhor aproxima a função considerada num certo ponto.

Assumindo que se pretende aproximar a função f num certo intervalo $[a, b]$ (ao qual pertencem os nós de interpolação), seja $\bar{x}$ um ponto genérico deste intervalo. Naturalmente, se $\bar{x}$ coincidir com um dos nós x_i teremos $e_n(\bar{x}) = e_n(x_i) = f(x_i) - P_n(x_i) = 0$.

Suponhamos que $\bar{x}$ não é nenhum dos nós. Para avaliar o erro de interpolação em $\bar{x}$, $e_n(\bar{x})$, vamos construir o polinómio P_{n+1} que interpola f em $x_0, x_1, \ldots, x_n, \bar{x}$. De acordo com a fórmula interpoladora de Newton, temos

$$P_{n+1}(x) = P_n(x) + f[x_0, x_1, ..., x_n, \bar{x}] \prod_{i=0}^{n} (x - x_i) \ . \tag{4.25}$$

Em particular,

$$P_{n+1}(\bar{x}) = P_n(\bar{x}) + f[x_0, x_1, ..., x_n, \bar{x}] \prod_{i=0}^{n} (\bar{x} - x_i) \ .$$

Dado que, por construção, $P_{n+1}(\bar{x}) = f(\bar{x})$, temos $e_n(\bar{x}) = P_{n+1}(\bar{x}) - P_n(\bar{x})$, e de (4.25) resulta

$$e_n(\bar{x}) = P_{n+1}(\bar{x}) - P_n(\bar{x}) = f[x_0, x_1, ..., x_n, \bar{x}] \prod_{i=0}^{n} (\bar{x} - x_i) \ . \tag{4.26}$$

Visto que $\bar{x}$ é um ponto genérico de $[a, b]$, a fórmula (4.26), pág. 203, pode ser utilizada para estimar o erro de interpolação em qualquer ponto deste intervalo.

A fórmula (4.26) não é facilmente aplicável, já que a estimativa do erro que ela proporciona depende de $f[x_0, x_1, ..., x_n, \bar{x}]$, grandeza que geralmente não é conhecida (aliás, em geral, nem sequer a função f é supostamente conhecida no ponto $\bar{x}$). Assim, para que a fórmula (4.26) possa ter alguma utilidade prática, é necessário relacionar as diferenças divididas de uma função f com as suas derivadas (assumindo que estas existem e podem ser calculadas).

4.1.7 Relação entre diferenças divididas e derivadas

No caso de $n = 1$ existe uma relação simples entre as diferenças divididas de uma função e a sua primeira derivada. De facto, se f for uma função continuamente diferenciável em $[x_0, x_1]$, de acordo com o teorema de Lagrange, pág. 32, existe pelo menos um ponto $\xi \in (x_0, x_1)$, tal que

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(\xi) \ . \tag{4.27}$$

Surge naturalmente a pergunta: será que existe uma relação semelhante entre as diferenças divididas de uma certa ordem k e a derivada de f da mesma ordem? A resposta a esta pergunta é positiva e é dada pelo teorema a seguir, que constitui uma generalização do referido teorema de Lagrange.

Teorema 4.1. Seja $f \in C^k([a,b])$, para $k \geq 1$, uma função dada e $x_0, x_1, ..., x_k$ um conjunto de $k+1$ pontos distintos do intervalo $[a,b]$. Existe pelo menos um ponto $\xi \in [a,b]$, tal que

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!} \ . \tag{4.28}$$

Demonstração. Seja

$$e_k(x) = f(x) - P_k(x)$$

o erro de interpolação de f por P_k, onde P_k representa o polinómio interpolador de f em $x_0, x_1, \ldots, x_k$. Por definição, temos

$$e_k(x_i) = 0, \qquad i = 0, 1, ..., k,$$

ou seja, a função erro $e_k(x)$ possui pelo menos $k+1$ zeros distintos em $[a,b]$. Além disso, $e_k(x)$ tem pelo menos k derivadas contínuas em $[a,b]$, segundo resulta das hipóteses do teorema.

Aplicando k vezes o teorema de Rolle, conclui-se que $e_k^{(k)}$ se anula, pelo menos, uma vez em $[a,b]$. Logo, existe $\xi \in [a,b]$, tal que $e_k^{(k)}(\xi) = 0$.

Mostremos que para o ponto ξ é válida a igualdade (4.28). Com efeito, pela definição de diferença dividida de ordem k, temos

$$0 = e_k^{(k)}(\xi) = f^{(k)}(\xi) - P_k^{(k)}(\xi) = f^{(k)}(\xi) - k!\, f[x_0, \ldots, x_k] \ . \tag{4.29}$$

Portanto,

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!}, \qquad \xi \in (a,b), \tag{4.30}$$

como se pretendia demonstrar. □

Fórmula teórica do erro de interpolação

Assumindo que no intervalo $[a,b]$, contendo um suporte de $n+1$ nós de interpolação, a função f é regular (pelo menos de classe $C^{n+1}([a,b])$), podemos concluir de (4.26) e (4.30) (fazendo $k = n$), que o erro de interpolação, para qualquer ponto $x \in [a,b]$, pode escrever-se na forma

$$\begin{aligned} e_n(x) &= f(x) - P_n(x) \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x - x_0)\,(x - x_1)\,\cdots\,(x - x_n) \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!}\,w_{n+1}(x), \end{aligned} \tag{4.31}$$

onde o polinómio w_{n+1}, de grau $n+1$, está associado aos nós de interpolação e, por isso, se designa habitualmente por *polinómio nodal*.

Na expressão de erro (4.31), o ponto $\xi = \xi(x)$ (dependente do ponto x) é geralmente desconhecido. Por isso, a expressão de erro anterior, embora de grande importância teórica, não é directamente aplicável quando se pretenda estimar o erro de interpolação num dado ponto do intervalo $[a, b]$ (erro de interpolação local), ou em qualquer ponto desse intervalo (erro de interpolação global). No entanto, a partir da fórmula teórica de erro poderemos obter majorações do respectivo erro absoluto, conforme se descreve no parágrafo a seguir.

4.1.8 Majoração do erro de interpolação

Da igualdade (4.30) resulta imediatamente que

$$|f[x_0, x_1, \ldots, x_k]| \leq \frac{1}{k!} \max_{x\in[a,b]} |f^{(k)}(x)| \ . \tag{4.32}$$

Combinando esta fórmula com (4.26), obtém-se a seguinte desigualdade fundamental,

$$\begin{aligned} |e_n(x)| &= |f[x_0, x_1, ..., x_n, \bar{x}]| \prod_{i=0}^{n} |x - x_i| \\ &\leq \frac{1}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)| \prod_{i=0}^{n} |x - x_i| \ . \end{aligned} \tag{4.33}$$

Exemplo 4.6. *Suponhamos que a função $f(x) = \cos(x)$ é aproximada no intervalo $[-1, 1]$ por um polinómio interpolador nos três nós $x_0 = -1$, $x_1 = 0$ e $x_2 = 1$.*

Verifiquemos que o erro máximo de interpolação em $[-1, 1]$, ocorre simetricamente relativamente à origem e perto das extremidades do intervalo em causa, conforme se ilustra na Figura 4.3.

(a) Determinar o polinómio interpolador $P_2(x)$.

(b) Determinar um majorante de $e_2(\bar{x})$ sendo $\bar{x} \in [-1, 1]$, ou seja, um majorante do erro de interpolação local.

(c) Determinar um majorante do erro máximo de interpolação no intervalo $[-1, 1]$, isto é, um majorante do erro de interpolação global.

(a) A fim de aplicar a fórmula interpoladora de Newton, comecemos por calcular

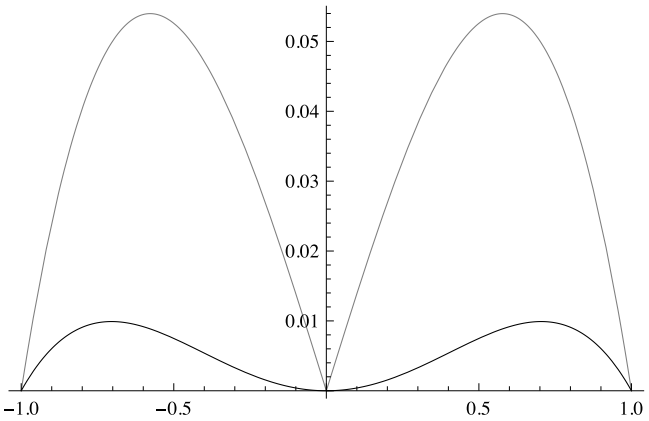


Figura 4.3: Para o Exemplo 4.6, o erro de interpolação absoluto de facto cometido está representado a traço espesso; a traço fino está representado o majorante do erro absoluto, dado pela fórmula (4.35).

as diferenças divididas de f,

$$\begin{aligned} f[x_0, x_1] &= \frac{\cos(x_1) - \cos(x_0)}{x_1 - x_0} = 1 - \cos(-1) \\ f[x_1, x_2] &= \frac{\cos(x_2) - \cos(x_1)}{x_2 - x_1} = \cos(1) - 1 \\ f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_1, x_2]}{x_2 - x_0} = \frac{2\cos(1) - 2}{2} = \cos(1) - 1 \ . \end{aligned}$$

Usando a fórmula (4.22), pág. 196, o polinómio interpolador é dado por,

$$\begin{aligned} P_2(x) &= f(-1) + f[-1, 0]\,(x+1) + f[-1, 0, 1]\,(x+1)\,x \\ &= \cos(-1) + (1 - \cos(-1))\,(x+1) + (\cos(1) - 1)\,(x+1)\,x \ . \end{aligned} \tag{4.34}$$

(b) De acordo com (4.33), o erro de interpolação localizado em $\bar{x}$ é majorado por

$$|e_2(\bar{x})| \leq \frac{1}{3!} \max_{x \in [-1,1]} |f^{(3)}(x)||\bar{x} + 1||\bar{x}||\bar{x} - 1| \ .$$

Além disso,

$$\max_{x \in [-1,1]} |f^{(3)}(x)| = \max_{x \in [-1,1]} |\sin(x)| = \sin(1).$$

Por conseguinte,

$$|e_2(\bar{x})| \leq \frac{\sin(1)}{3!} |\bar{x} + 1||\bar{x}||\bar{x} - 1| \ . \tag{4.35}$$

(c) Pretende-se majorar $E = \max_{\bar{x} \in [-1,1]} |e_2(\bar{x})|$. Para isso, baseando-nos na resposta anterior, basta obter

$$\max_{\bar{x} \in [-1,1]} |w_3(\bar{x})|,$$

onde

$$w_3(x) = x\,(x-1)\,(x+1) = x^3 - x\ .$$

Para determinar os pontos de extremo de $w_3(x)$, resolve-se a equação

$$w_3'(x) = 3\,x^2 - 1 = 0,$$

a qual tem como raízes reais $\alpha_1 = -\dfrac{1}{\sqrt{3}}$ e $\alpha_2 = \dfrac{1}{\sqrt{3}} \simeq 0.58$. É fácil verificar que a primeira destas raízes corresponde a um máximo local de w_3, enquanto a segunda refere-se a um mínimo local. Por outro lado, sendo w_3 uma função ímpar, facilmente se deduz que o mínimo local é o simétrico do máximo local. Assim,

$$\max_{\bar{x}\in[-1,1]} |w_3(\bar{x})| = |w_3(\alpha_1)| = |w_3(\alpha_2)| = |\alpha_2\,(\alpha_2-1)\,(\alpha_2+1)| = \frac{2}{3\sqrt{3}}\ . \quad (4.36)$$

Finalmente, combinando (4.35) com (4.36), obtém-se

$$E = \max_{\bar{x}\in[-1,1]} |e_2(\bar{x})| \le \frac{\sin(1)}{3!} \max_{\bar{x}\in[-1,1]} |w_3(\bar{x})| = \frac{\sin(1)}{3!}\frac{2}{3\sqrt{3}} \approx 0.054\ . \quad (4.37)$$

♦

4.1.9 O exemplo de Runge *

Polinómios interpoladores construídos a partir de um suporte com nós de interpolação *equidistantes* são susceptíveis de produzir oscilações de grande amplitude próximo dos extremos do intervalo de interpolação, oscilações tanto maiores quanto maior for o número de nós de interpolação. Esta característica indesejável é conhecida como "fenómeno de Runge"[5].

No célebre exemplo de Runge, a função a aproximar é

$$f(x) = \frac{1}{1+25\,x^2}, \qquad -1 \le x \le 1\ . \quad (4.38)$$

Trata-se de uma função par e continuamente diferenciável para qualquer ordem, ou seja de classe $C^\infty([-1,1]$.

Fixado $n \ge 1$, considerem-se os $n+1$ nós equidistantes,

$$x_0 = -1 + i\,h, \quad \text{com} \quad h = \frac{2}{n}, \quad \text{para} \quad i = 0:n\ .$$

Para esta *malha de interpolação* uniforme, é natural perguntar se à medida que se aumentam o número de nós da malha, o respectivo polinómio interpolador se aproxima ou não da função f.

[5]Carl David Tolmé Runge, 1856–1927, matemático e físico alemão.

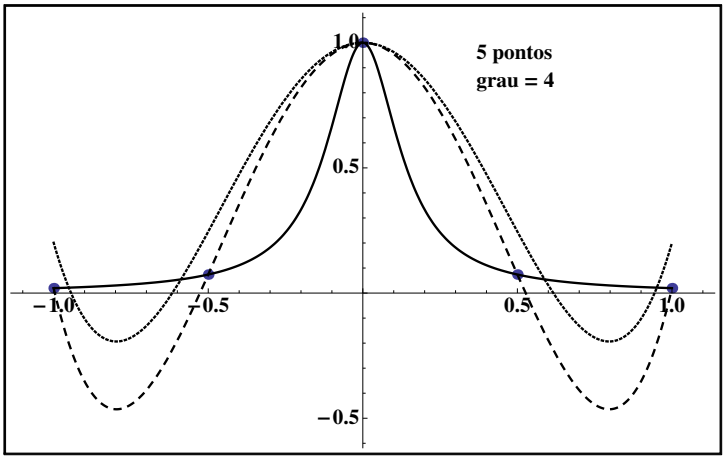


Figura 4.4: Exemplo de Runge para $n = 5$. A função (4.38) está representada a traço cheio.

Mais formalmente, pretende-se saber se a distância entre f e o polinómio interpolador $P_n(x)$ (distância essa medida na norma a seguir) decresce com n, no sentido seguinte:

$$\lim_{n\to\infty} ||f - P_n||_\infty = \lim_{n\to\infty} (max_{-1\le x\le 1}|f(x) - P_n(x)|) = 0 \ .$$

As figuras 4.4 e 4.5 ilustram ser negativa a resposta a essa questão, porquanto contrariamente ao que a intuição nos poderia levar a pensar, o polinómio interpolador P_n não se aproxima da função f à medida que n aumenta.

Na Figura 4.4 evidencia-se esse facto mostrando a tracejado grosso o polinómio interpolador $P_5(x)$ e na Figura 4.5 o polinómio interpolador $P_{15}(x)$. Este último apresenta enormes oscilações próximo dos extremos do intervalo $[-1, 1]$, logo afasta-se da função (a traço cheio) em vez de se aproximar. Pelo contrário, nas referidas figuras surge ainda a tracejado fino, respectivamente um polinómio interpolador de grau 5 e de grau 15, usando nós de interpolação não igualmente espaçados. Esses dois polinómios interpoladores não têm o inconveniente anteriormente apontado, sendo que o polinómio de grau 15 aproxima melhor a função em todo o intervalo do que o polinómio de grau 5.

Que malha de interpolação é usada por tais polinómios "bem comportados"?

Fixado n, a malha de interpolação referida é constituída pelos *zeros* do chamado polinómio de Chebyshev[6] de grau n. No presente exemplo, para $n = 5$ (Figura 4.4), a malha de interpolação é constituída pelos zeros do polinómio de Chebyshev T_5,

$$\begin{array}{ll} T_5(t) & = 5\,t - 20\,t^3 + 16\,t^5 \\ \text{zeros} & \to -0.951057, -0.587785, 0., 0.587785, 0.951057 \ . \end{array} \tag{4.39}$$

[6]Pafnuty Lvovich Chebyshev, 1821 -1894, matemático russo.

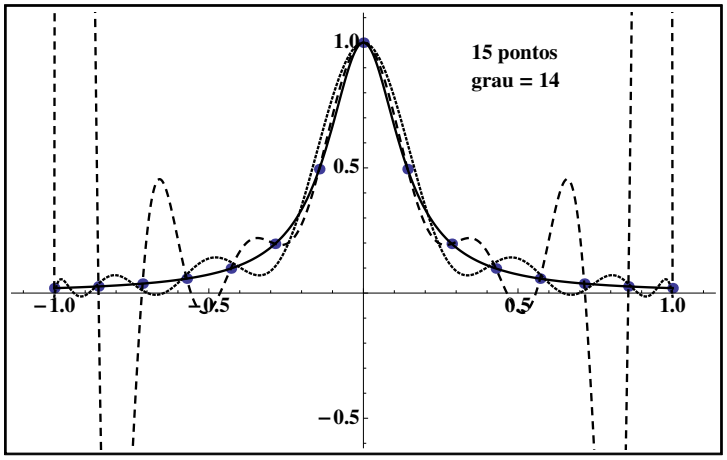


Figura 4.5: Exemplo de Runge para $n = 15$.

e para $n = 15$ (Figura 4.5), a malha é constituída pelos zeros do polinómio de Chebyshev T_{15},

$$\begin{array}{ll} T_{15}(t) & = -15\,t + 560\,t^3 - 6048\,t^5 + 28800\,t^7 - 70400\,t^9 + 92160\,t^{11} - \\ & -61440\,t^{13} + 16384\,t^{15} \\ \text{zeros} & \to -0.994522, -0.951057, -0.866025, -0.743145, -0.587785, \\ & -0.406737, -0.207912, 0., 0.207912, 0.406737, 0.587785, \\ & 0.743145, 0.866025, 0.951057, 0.994522\ . \end{array} \tag{4.40}$$

Os zeros anteriores são aproximações obtidas por arredondamento simétrico.

Os polinómios de Chebyshev constituem uma importante família de funções polinomiais com aplicação em diversos ramos da matemática. Para $n \geq 0$, estes polinómios podem ser definidos pela expressão

$$T_n(t) = \cos(n\,\arccos t), \qquad t \in [-1, 1], \tag{4.41}$$

donde

$$T_n\left(\cos(\theta)\right) = \cos(n\,\theta), \qquad \theta \in [0, \pi], \tag{4.42}$$

Os polinómios de Chebyshev podem obter-se recursivamente. De facto, atendendo à expressão trigonométrica

$$\cos((n+1)\,\theta) + \cos((n-1)\,\theta) = 2\,\cos(\theta)\,\cos(n\,\theta), \quad \forall n \geq 1$$

resulta

$$\cos((n+1)\,\theta) = 2\,\cos(\theta)\,\cos(n\,\theta) - \cos((n-1)\,\theta), \quad \forall n \geq 1\ .$$

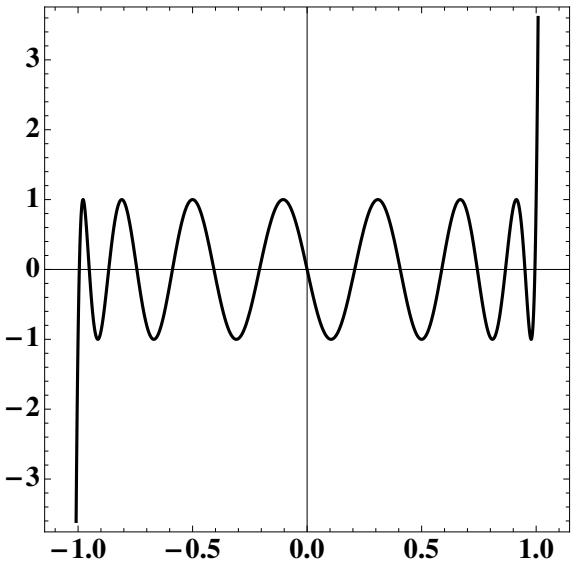


Figura 4.6: Polinómio de Chebyshev $T_{15}(t)$.

Da fórmula de recorrência anterior, obtém-se

$$\begin{array}{ll} \cos(2\,\theta) & = 2\,\cos^2(\theta) - 1 \\ \cos(3\,\theta) & = 2\,\cos(\theta)\,\cos(2\,\theta) - \cos(\theta) = 4\,\cos^3(\theta) - 3\,\cos(\theta) \\ & \vdots \end{array}$$

Fazendo

$$t = \cos(\theta) \quad \Longrightarrow \quad \theta = \arccos(t),$$

verificamos assim que a expressão (4.42) define de facto um polinómio de grau n.

A recursividade dos polinómios de Chebyshev, anteriormente esboçada, traduz-se nas expressões

$$\begin{array}{lll} T_0(t) & = 1 & \\ T_1(t) & = t & \\ T_{k+1}(t) & = 2\,t\,T_k(t) - T_{k-1}(t), \qquad & k = 1, 2, \ldots, \end{array} \tag{4.43}$$

pelo que os primeiros polinómios de Chebyshev, de grau ≥ 2, são os seguintes:

$$\begin{array}{ll} T_2(t) & = 2\,t^2 - 1 \\ T_3(t) & = 4\,t^3 - 3\,t \\ T_4(t) & = 8\,t^4 - 8\,t^2 + 1 \\ & \vdots \end{array}$$

Zeros dos polinómios de Chebyshev

Dado que para $n \geq 1$, se tem

$$\cos(n\,\theta) = 0 \iff n\,\theta = \pi/2 + k\,\pi = \pi/2\,(1+2\,k) \iff \theta = \frac{\pi}{2}\,\frac{1+2\,k}{n},$$

os zeros do polinómio de Chebyshev $T_n(t)$, são os seguintes pontos do intervalo $(-1, 1)$,

$$t_i = \cos(\theta_i) = \cos\left(\frac{1+2\,i}{n} \times \frac{\pi}{2}\right), \qquad i = 0 : (n-1) \tag{4.44}$$

Um suporte de interpolação que use os nós t_i dir-se-á um suporte de Chebyshev.

Convida-se o leitor a verificar que, respectivamente para $n = 5$ e $n = 15$, o suporte de Chebyshev que anteriomente usámos para obter as Figuras 4.4 e 4.5 é constituído pelos pontos indicados em (4.39) e (4.40). Na Figura 4.6 está representado o polinómio de Chebyshev de grau 15, $T_{15}(t)$, com $t \in [-1.1, 1.1]$. Note-se que $T_{15}(t)$ toma valores entre -1 e 1, no intervalo $[-1, 1]$, como seria de esperar.

No Exercício 4.2, pág. 235, é ilustrada a vantagem que existe na escolha de um suporte de Chebyshev, tendo em vista minorar o erro de interpolação num intervalo.

Entre outras aplicações, a interpolação de Chebyshev desempenha um papel fundamental no cálculo de raízes de equações $f(x) = 0$. A partir dos coeficientes do polinómio interpolar de Chebyshev é construída a sua *matriz companheira*, cujos valores próprios são os zeros da função dada f. Sobre este interessante algoritmo, cruzando ideias da teoria da aproximação de funções com a geometria algébrica e a álgebra linear, aconselha-se a leitura de J. Boyd [8, 7].

4.1.10 Fórmulas baricêntricas do polinómio interpolador de Lagrange *

O polinómio interpolador de Lagrange pode ser rescrito utilizando fórmulas computacionalmente mais eficientes do que a fórmula clássica que anteriormente discutimos (ver página 189). Estas fórmulas recebem a designação de fórmulas *baricêntricas* do polinómio de Lagrange.

Fixado $n \geq 0$, e dados $n+1$ nós distintos x_j, bem como os correspondentes valores f_j de uma tabela, para $j = 0, \ldots, n$, recorde-se que a fórmula interpeladora de Lagrange se escreve

$$p_n(x) = \sum_{j=0}^{n} l_j(x)\, f_j, \quad \text{onde} \quad l_j(x) = \frac{\prod_{k=0}^{n}(x - x_k)}{\prod_{k=0}^{n}(x_j - x_k)}\ . \tag{4.45}$$

Fixado um valor do argumento x e usando uma forma conveniente para a expressão de $p_n(x)$, as fórmulas baricêntricas a seguir referidas permitem calcular $p_n(x)$ mediante $\mathcal{O}(n^2)$ operações elementares, tal como a forma de Newton que já conhecemos (ver parágrafo 4.1.4, pág. 195). Dado que as quantidades envolvendo $\mathcal{O}(n^2)$ operações nas fórmulas baricêntricas não dependem dos valores tabelados f_j, tais fórmulas podem ser úteis para obter o polinómio interpolador de funções distintas definidas no mesmo conjunto de nós $x_0, \ldots, x_n$. Pelo contrário, como sabemos, a fórmula interpeladora de Newton exige que se efectue o cálculo da tabela de diferenças divididas para cada uma das funções que se considere.

Comecemos por definir o *polinómio nodal*, de grau $n+1$,

$$\Omega(x) = (x - x_0)\,(x - x_1)\,\cdots\,(x - x_n), \tag{4.46}$$

e para cada nó x_j, o *peso baricêntrico* de x_j,

$$\omega_j = \frac{1}{\prod_{k\neq j}(x_j - x_k)}, \quad j = 0, \ldots, n\ . \tag{4.47}$$

Usando a notação anterior, cada elemento $l_j(x)$ da base de Lagrange considerado em (4.45) passa a escrever-se na forma

$$l_j(x) = \frac{\omega_j}{(x - x_j)}\,\Omega(x), \quad j = 0, \ldots, n\ . \tag{4.48}$$

Note-se que os pesos ω_j não dependem do ponto de interpolação x. O conjunto destes pesos, pode ser calculado mediante $\mathcal{O}(n^2)$ operações elementares. Para se obter $p_n(x)$ são ainda necessárias mais $\mathcal{O}(n)$ operações, pelo que o valor do polinómio interpolador num ponto por meio das fórmulas baricêntricas dadas a seguir pode ser obtido efectuando $\mathcal{O}(n^2)$ operações.

Levando em consideração (4.47) e (4.48) e atendendo a que para $j = 0, \ldots, n$ os termos da soma em (4.45) contêm o polinómio nodal $\Omega(x)$ (que não depende de j), resulta

$$p_n(x) = \Omega(x) \sum_{j=0}^{n} \frac{\omega_j}{(x - x_j)}\, f_j, \tag{4.49}$$

expressão que recebe a designação de *primeira fórmula baricêntrica* do polinómio interpolador de Lagrange.

Notando que o polinómio interpolador da função $f(x) = 1$ é a própria função, de (4.49) resulta

$$1 = \Omega(x) \sum_{j=0}^{n} \frac{\omega_j}{(x - x_j)}\ . \tag{4.50}$$

Assim, o quociente $p_n(x)/1$, levando em consideração as igualdades (4.50) e (4.49), passa a ser

$$p_n(x) = \frac{\sum_{j=0}^{n} \dfrac{\omega_j\, f_j}{(x - x_j)}}{\sum_{j=0}^{n} \dfrac{\omega_j}{(x - x_j)}}, \tag{4.51}$$

onde o peso ω_j é definido por (4.47). A fórmula (4.51) é conhecida pela designação de *segunda fórmula baricêntrica* do polinómio de Lagrange. Os autores Berrut e Trefethen apresentam em [3] uma discussão interessante sobre aplicações das fórmulas baricêntricas de Lagrange.

Para finalizarmos esta secção, refira-se que nas aplicações é frequentemente utilizado outro tipo de interpolação que não o polinomial. Nomeadamente a interpolação *racional* (cujas funções aproximantes são quocientes de polinómios). Sobre esta matéria convida-se o leitor a ler o interessante artigo (em francês) [4].

4.2 Interpolação polinomial bivariada *

Interpolação espacial ou multivariada é utilizada em diversas diversas áreas como, por exemplo, computação gráfica, *quadratura* e *cubatura* numéricas e na resolução de equações diferenciais.

Fixados nós distintos num intervalo real $[a, b]$, sabemos que o problema de interpolação polinomial tem solução única (ver 4.1.1, pág. 184). Quando passamos a pontos distintos no espaço linear $\mathbb{R}^d$, para $d \geq 2$, um problema interpolatório pode não ter solução ou ter uma infinidade delas.

Para ilustrarmos o problema, damos exemplos a seguir de construção de polinómios de duas variáveis em $\mathbb{R}^2$ e valores em $\mathbb{R}$, nos quais a configuração de pontos dados no plano pode levar a um problema de interpolação impossível ou indeterminado. Mostramos depois que numa *malha rectangular* o problema de interpolação bivariada tem solução única, assim como em certas *malhas triangulares*. Suportes de interpolação bivariada destes dois tipos ocorrem frequentemente nas aplicações.

No parágrafo 4.2.2 é introduzido o polinómio interpolador, a duas variáveis, na forma de Lagrange. O mesmo polinómio usando certas diferenças divididas generalizadas constitui uma extensão do correspondente polinómio de Newton (ver pág. 195). Ao leitor interessado em aprofundar o tema recomenda-se a leitura da obra [30].

4.2.1 Existência e unicidade de polinómio interpolador

Dados quatro pontos A_0, A_1, A_2 e A_3, dispostos no plano como se indica na Figura 4.7, com valores $f_{i,j} = f(x_i, y_j) \in \mathbb{R}$ (ver tabela 4.3), pretende-se determinar os

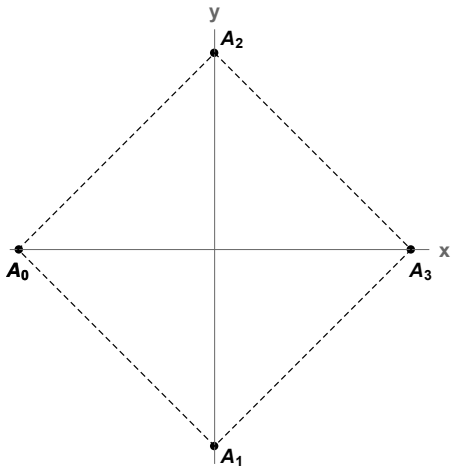


Figura 4.7: Uma configuração quadrada para quatro pontos do plano.

coeficientes do polinómio modelo

$$p(x,y) = a_0 + a_1\,x + a_2\,y + a_3\,x\,y,$$

de modo que este interpole os dados, ou seja que, $p(x_i, y_j) = f(x_i, x_j)$, para $i, j = 0, 1$. O polinómio modelo, de grau ≤ 2 (se $a_3 \neq 0$ dizemos tratar-se de polinómio do segundo grau já que a soma das potências das suas variáveis é ≤ 2) é uma combinação linear das funções polinomiais de base

$$\begin{array}{ll} \phi_0(x,y) = 1, & \text{grau } 0 \\ \phi_1(x,y) = x, & \text{grau } 1 \\ \phi_2(x,y) = y, & \text{grau } 1 \\ \phi_3(x,y) = x\,y, & \text{grau } 2\ . \end{array}$$

Este problema de interpolação pode assim traduzir-se no sistema $M\,a = f$,

$$M\,a = f \quad \Leftrightarrow \quad \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(-1,0) \\ f(0,-1) \\ f(0,1) \\ f(1,0) \end{bmatrix}.$$

Note-se que as colunas da matriz M se obtém como resultado da aplicação respectivamente de $\phi_0(x,y)$, $\phi_1(x,y)$, $\phi_2(x,y)$ e $\phi_3(x,y)$ aos 4 pontos do plano A_0, A_1, A_2 e A_3. Se designarmos essas colunas por ϕ_0 a ϕ_3, vemos que a última coluna é constituida por entradas nulas. Isto significa que o sistema é singular e portanto ou é impossível ou indeterminado.

Para uma configuração triangular de quatro pontos do plano, B_0 a B_3 (ver Figura 4.8 e tabela 4.4), o mesmo modelo polinomial de $p(x,y)$ traduz-se no novo sistema linear $M\,a = f$, onde

Pontos	Valor
$A_0 = (-1, 0)$	$f(-1, 0)$
$A_1 = (0, -1)$	$f(0, -1)$
$A_2 = (0, 1)$	$f(0, -1)$
$A_3 = (1, 0)$	$f(0, -1)$

Tabela 4.3: Quatro pontos do plano dispostos num quadrado.

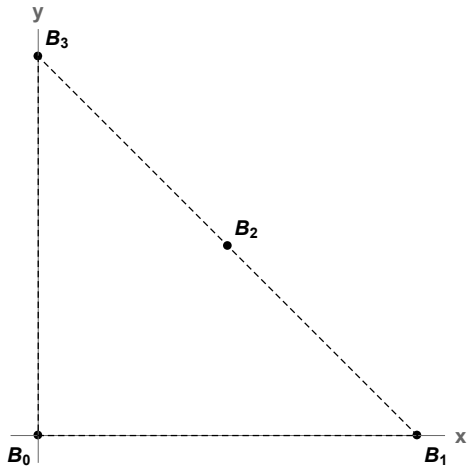


Figura 4.8: Uma configuração triangular para quatro pontos do plano.

$$M\,a = f \quad \Leftrightarrow \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(0,0) \\ f(2,0) \\ f(1,1) \\ f(0,2) \end{bmatrix}.$$

Pontos	Valor
$B_0 = (0, 0)$	$f(0, 0)$
$B_1 = (2, 0)$	$f(2, 0)$
$B_2 = (1, 1)$	$f(1, 1)$
$B_3 = (0, 2)$	$f(0, 2)$

Tabela 4.4: Quatro pontos do plano dispostos num triângulo.

Tem-se que $det(M) = -4 \neq 0$ (a solução é única) e

$$\begin{aligned} a_0 &= f(0,0) \\ a_1 &= (f(2,0) - f(0,0))/2 \\ a_2 &= (f(0,2) - f(0,0))/2 \\ a_3 &= -(f(0,2) - 2\,f(1,1) + f(2,0))/2\ . \end{aligned}$$

Assim, constatámos experimentalmente que a determinação de um polinómio interpolador depende da disposição espacial dos nós de interpolação usados. No primeiro caso (4 nós A_i do plano dispostos como vértices de um determinado quadrado) somos conduzidos a um problema interpolatório impossível ou indeterminado; no segundo caso (4 nós B_i dispostos num certo triângulo) o problema interpolatório tem solução única.

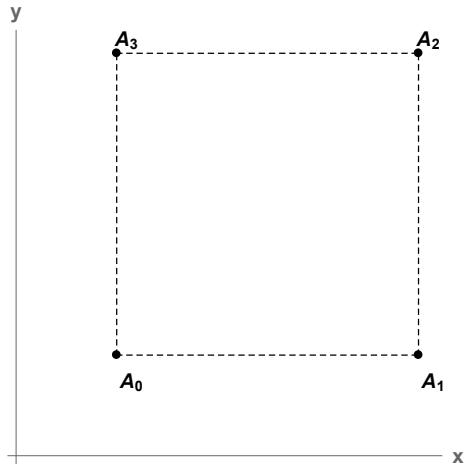


Figura 4.9: Uma configuração rectangular para quatro pontos do plano.

Para um dado número natural n, considerando $N = \binom{n+2}{2}$ pontos distintos do plano, em disposição triangular, de coordenadas cartesianas (x_i, y_j), tal que $i, j \geq 0$, $i + j \leq n$, pode provar-se (ver [30], p. 178) que o respectivo problema de interpolação polinomial tem solução única.

Malha rectangular

Para $X = \{x_0, x_1\}$, com $x_0 \neq x_1$ e $Y = \{y_0, y_1\}$, onde $y_0 \neq y_1$, uma malha rectangular genérica de 4 pontos do plano, definida pelo produto cartesiano $X \times Y$, é mostrada na Figura 4.9. Reutilizando o modelo interpolatório anteriormente considerado, $p(x, y) = a_0 + a_1\,x + a_2\,y + a_3\,x\,y$, obtém-se o correspondente sistema linear $M\,a = f$ da forma,

$$M\,a = f \quad \Leftrightarrow \quad \begin{bmatrix} 1 & x_0 & y_0 & x_0\,y_0 \\ 1 & x_1 & y_0 & x_1\,y_0 \\ 1 & x_1 & y_1 & x_1\,y_1 \\ 1 & x_0 & y_1 & x_0\,y_1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(A_0) \\ f(A_1) \\ f(A_2) \\ f(A_3) \end{bmatrix}.$$

A matriz M pode factorizar-se no produto $L\,U$, onde L é triangular inferior e U triangular superior, sendo

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \; U = \begin{bmatrix} 1 & x_0 & y_0 & x_0\,y_0 \\ 0 & x_1 - x_0 & 0 & (x_1 - x_0)\,y_0 \\ 0 & 0 & y_1 - y_0 & x_1\,(y_1 - y_0) \\ 0 & 0 & 0 & (x_1 - x_0)\,(y_0 - y_1) \end{bmatrix}.$$

Dado que $det(M) = det(L)\,det(U) = (x_1 - x_0)\,(y_1 - y_0)\,(x_1 - x_0)\,(y_0 - y_1)$, uma vez que admitimos que $x_i \neq x_j$ e $y_i \neq y_j$, tem-se que $det(M) \neq 0$, o mesmo é dizer que o problema interpolatório considerado tem solução única.

De modo análogo, pode mostrar-se que para uma malha rectangular de $(m+1) \times (n+1)$ nós, resultando do produto cartesiano $X \times Y$, onde $X = (x_0, x_1, \ldots, x_m)$, $Y = (y_0, y_1, \ldots, y_n)$, com $x_i \neq x_j$ e $y_i \neq y_j$, $\quad (i \neq j)$, e dados os respectivos valores $f_{i,j} = f(x_i, y_j)$, para $i = 0, \ldots, m$ e $j = 0, \ldots, n$, existe um só polinómio interpolador de grau apropriado.

No parágrafo 4.2.2 adiante mostram-se exemplos de construção de polinómios interpoladores em malhas rectangulares $X \times Y$, de $(m+1) \times (n+1)$ pontos do plano, usando as bases clássicas de Lagrange, respectivamente nas variáveis x e y,

Exercício 4.1.

a) Confirme que para a configuração triangular dada na Figura 4.10, onde se apresentam 4 pontos denotados como $B_0 = (x_0, y_0)$, $B_1 = (x_1, y_1)$, $B_2 = (x_2, y_0)$ e $B_3 = (x_0, y_2)$, a matriz M associada a estes nós de $\mathbb{R}^2$ tem a forma

$$M = \begin{bmatrix} 1 & x_0 & y_0 & x_0\,y_0 \\ 1 & x_0 + t\,(x_2 - x_0) & y_2 - t\,(y_2 - y_0) & (x_0 + t(x_2 - x_0))\,(y_2 - t(y_2 - y_0)) \\ 1 & x_2 & y_0 & x_2\,y_0 \\ 1 & x_0 & y_2 & x_0\,y_2 \end{bmatrix},$$

assumindo que o ponto B_1 pode ocupar uma qualquer posição no segmento $\overline{B_2\,B_3}$ (exceptuando os extremos), isto é, para um valor do parâmetro $t \in (0, 1)$ as coordenadas do ponto B_1 são dadas por

$$\begin{cases} x_1 = x_0 + t\,(x_2 - x_0) \\ y_1 = y_2 + t\,(y_0 - y_2) \end{cases}.$$

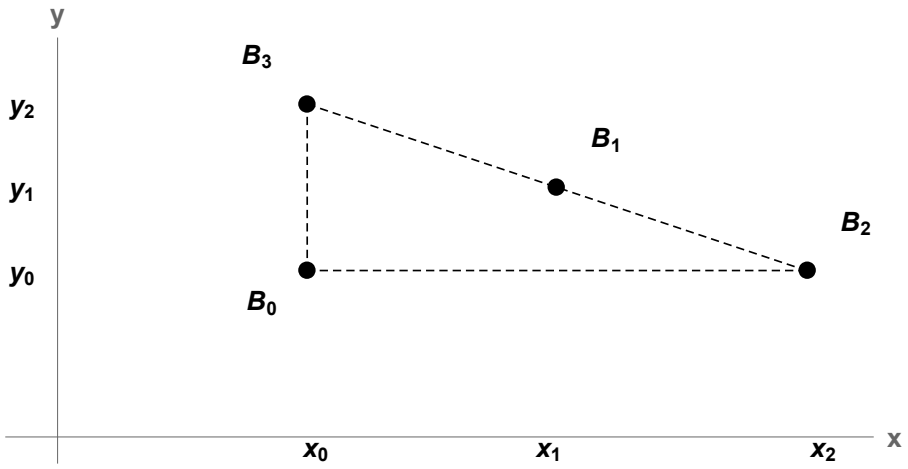


Figura 4.10: Uma configuração triangular para quatro pontos do plano.

Conclua que

$$det(M) = (1-t)\,t\,(x_2 - x_0)^2\,(y_2 - y_0)^2 \neq 0,$$

e que por conseguinte o modelo interpolatório $p(x,y) = a_0 + a_1 x + a_2\,y + a_3\,x\,y$ *tem solução única para configurações triangulares como a considerada.*

b) Se em vez de adoptar os nós de interpolação ordenados como na Figura 4.10, considerar quatro pontos $C_0 = (x_0, y_0)$, $C_1 = (x_2, y_0)$, $C_2 = (x_1, y_1)$ *e* $C_3 = (x_0, y_2)$, *qual é a respectiva matriz* M*? Conclua que o polinómio interpolador não depende da ordem estabelecida para os 4 nós de interpolação dispostos triangularmente.*

c) Para os valores inscritos na Tabela 4.5 mostre que o respectivo polinómio interpolador é $p(x,y) = 1$.
Sendo dada uma função $f(x,y)$, *integrável e com valores na região triangular considerada, o integral respectivo pode ser aproximado pelo integral do polinómio interpolador dos valores tabelados. Supondo que uma função* f *assume os valores que constam da referida tabela, mostre que*

$$\int_0^1 \int_{1/3}^{2/3} f(x,y)\,dxdy \simeq \int_0^1 \int_{1/3}^{2/3} dxdy = 1/3\ .$$

4.2.2 Polinómio interpolador na base de Lagrange

A fim de se generalizar o polinómio interpolador de Lagrange numa variável ao caso bidimensional, são introduzidas a seguir funções $a(x,y)$ e $b(x,y)$ que fazem intervir a base de Lagrange clássica, respectivamente na variável x e e na

$y_i \rightarrow$ $x_i \downarrow$	$1/3$	$1/2$	$2/3$
0	1	$*$	1
$1/2$	$*$	1	$*$
1	1	$*$	$*$

Tabela 4.5: $f(x_i, y_i) = 1$ para 4 pontos de suporte triangular.

variável y. Tais funções recebem a designação de *funções de mistura*[7], uma vez que elas permitem naturalmente construir um polinómio interpolador bivariado misturando convenientemente a base de Lagrange respectivamente nas variáveis x e y.

Para ilustrarmos a construção de um polinómio interpolador definido numa malha rectangular de $N = (m+1)\times(n+1)$ pontos distintos do plano, considere-se $m = 2$ e $n = 1$, sendo dados

$$X = \{x_0, x_1, x_2\}, \qquad Y = \{y_0, y_1\},$$

onde se conhecem valores $f_{i,j}$ nos $N = 6$ pontos de $X \times Y \subset \mathbb{R}^2$. Ao suporte X associamos a base de Lagrange clássica

$$< L_0(x), L_1(x), L_2(x) >,$$

onde $L_i(x) = \prod_{i,j\geq 0, i\neq j}^{m} (x - x_j)/(x_i - x_j)$, para $i = 0, 1, 2$ são polinómios de grau $m = 2$ na variável x. Ao suporte Y associamos a base de Lagrange

$$< M_0(y), M_1(y) >,$$

onde $M_j(y) = \prod_{i,j\geq 0, i\neq j}^{n} (y - y_j)/(y_i - y_j)$, para $j = 0, 1$ são, neste caso, polinómios de grau $n = 1$ na variável y.

Vejamos que é possível combinar os polinómios $L_i(x)$, com os polinómios $M_j(y)$ de modo a resultar um polinómio interpolador na malha rectangular considerada. As funções $a_f(x, y)$ e $b(x, y)$ a seguir definidas dizem-se *funções de mistura* uma vez que nos permitem obter o polinómio interpolador $p(x, y)$ misturando os polinómios de Lagrange, respectivamente nas variáveis x e y.

Seja

$$a_f(x, y) = f(x_0, y)\, L_0(x) + f(x_1, y)\, L_1(x) + f(x_2, y)\, L_2(x)\ . \tag{4.52}$$

Como $L_i(x_i) = 1$ e $L_i(x_j) = 0$ para $i \neq j$, tem-se

$$\begin{aligned} a_f(x_0, y) &= f(x_0, y) \\ a_f(x_1, y) &= f(x_1, y) \\ a_f(x_2, y) &= f(x_2, y)\ . \end{aligned}$$

[7]Na literatura anglo-saxónica "blending functions".

Definindo agora

$$b(x,y) = f(x,y_0)\,M_0(y) + f(x,y_1)\,M_1(y), \tag{4.53}$$

Como $M_i(y_i) = 1$ e $M_i(y_j) = 0$ para $i \neq j$, resulta

$$\begin{aligned} b(x,y_0) &= f(x,y_0) \\ b(x,y_1) &= f(x,y_1)\ . \end{aligned}$$

A mistura das referidas bases de Lagrange resulta de na expressão (4.52) se substituir f pela função $b(x,y)$, i.e., compor a função a_f com a função b, obtendo-se

$$p(x,y) = b(x_0,y)\,L_0(x) + b(x_1,y)\,L_1(x) + b(x_2,y)\,L_2(x)\ .$$

Ou seja,

$$\begin{aligned} p(x,y) \quad &= (f(x_0,y_0)\,M_0(y) + f(x_0,y_1)\,M_1(y))\ L_0(x) + \\ &+ (f(x_1,y_0)\,M_0(y) + f(x_1,y_1)\,M_1(y))\ L_1(x) + \\ &+ (f(x_2,y_0)\,M_0(y) + f(x_2,y_1)\,M_1(y))\ L_2(x), \end{aligned} \tag{4.54}$$

donde se conclui imediatamente que $p(x_i,y_j) = f_{i,j}$, para $i = 0,1,2,\ j = 0,1$, o que significa que o polinómio (4.54) é interpolador dos dados. Uma vez que o polinómio interpolador em malha rectangular sabermos ser único, a expressão anterior de $p(x,y)$ corresponde ao polinómio pretendido.

De modo análogo se conclui que para uma malha rectangular geral com $N = (m+1)\times(n+1)$ pontos distintos, de suportes $X = \{x_0,x_1,\ldots,x_m\}$, $Y = \{y_0,y_1,\ldots,y_n\}$, onde são dados valores $f_{i,j}$ em $X\times Y \subset \mathbb{R}^2$, o polinómio

$$p(x,y) = \sum_{i=0}^{m}\left(\sum_{j=0}^{n} f(x_i,y_j)\,M_j(y)\right) L_i(x) = \sum_{j=0}^{n}\left(\sum_{i=0}^{m} f(x_i,y_j)\,L_i(x)\right) M_j(y) \tag{4.55}$$

é interpolador.

Exemplo 4.7. *Sejam $m = 2$, $n = 1$, e os suportes $X = \{-3,0,2\}$, $Y = \{-1,1\}$. Na tabela 4.6 são dados valores $f_{i,j}$ da função $f(x,y) = x\,3^y - y\,2^{x+y}$ nos $N = 6$ pontos definidos por $X\times Y$.*
A base de Lagrange associada ao suporte X é constituida pelos seguintes polinómios de grau 2,

$$\begin{aligned} L_0(x) &= (x\,(x-2))\,/15 \\ L_1(x) &= -\left((x+3)\,(x-2)\right)/6 \\ L_2(x) &= -\left((x+3)\,x\right)/10, \end{aligned}$$

enquanto que a base de Lagrange associada ao suporte Y contém os polinómios de grau 1,

$$\begin{aligned} M_0(y) &= -\,(y-1)\,/2 \\ M_1(y) &= -\,(y+1)\,/2\ . \end{aligned}$$

$y_i \rightarrow$ $x_i \downarrow$	-1	1
-3	$-15/16$	$-37/4$
0	$1/2$	-2
2	$8/3$	-2

Tabela 4.6: Malha rectangular com $N = 6$ pontos.

Atendendo à expressão (4.55)*, o polinómio interpolador dos dados escreve-se*

$$\begin{aligned} p(x,y) &= (-15/16\, L_0(x) + 1/2\, L_1(x) + 8/3\, L_2(x))\ M_0(y) + \\ &\quad + (-37/4\, L_0(x) - 2\, L_1(x) - 2\, L_2(x))\ M_1(y) \\ &= 1/480\ (-360 + 434\, x - 87\, x^2 - 600\, y + 30\, xy - 145\, x^2\, y)\ . \end{aligned}$$

Assim, se $[V]_{i=0,2}^{j=0,1}$ *designar a matriz das* observações $f_{i,j}$*, e* $\mathcal{L}$*,* $\mathcal{M}$ *designarem vectores coluna contendo respectivamente os elementos da base de Lagrange de* X *e de* Y*, o polinómio interpolador pode ser escrito na forma matricial*

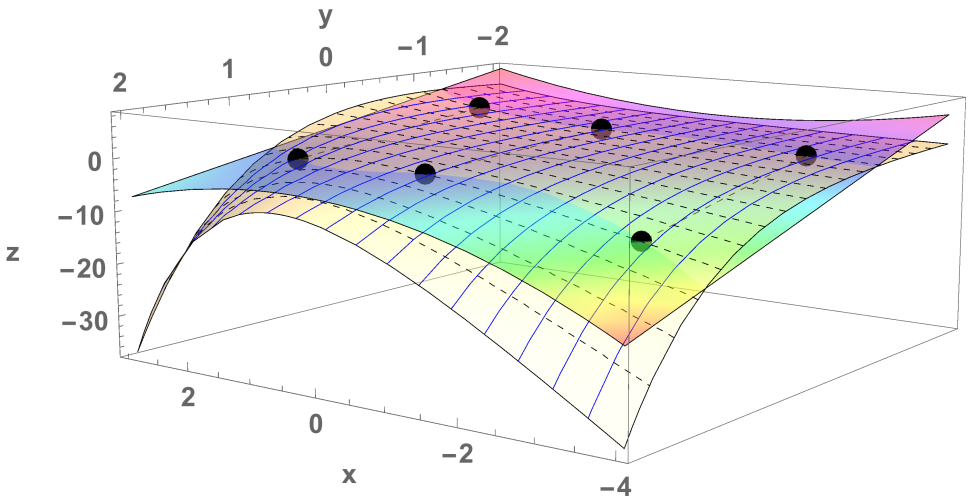


Figura 4.11: Polinómio interpolador para uma malha rectangular de $N = 6$ pontos em $[-3, 2] \times [-1, 1]$.

$$\begin{aligned} p(x,y) &= \mathcal{M}^T\, V^T\, \mathcal{L} \\ &= \begin{bmatrix} M_0(y) & M_1(y) \end{bmatrix} \begin{bmatrix} -15/16 & 1/2 & 8/3 \\ -37/4 & -2 & -2 \end{bmatrix} \begin{bmatrix} L_0(x) \\ L_1(x) \\ L_2(x) \end{bmatrix}. \end{aligned}$$

Na Figura 4.11 mostra-se a superfície definida pela função $f(x, y)$*, conjuntamente com os pontos da tabela 4.6, bem como a superfície que se lhe sobrepõe, definida pelo polinómio interpolador* $p(x, y)$ *anteriormente calculado.*

Se se considerar uma malha quadrada de $N = 25$ pontos definida a partir dos suportes $X = Y = \{-2, -1, 0, 1, 2\}$, para a mesma função $f(x, y)$ anterior, pode verificar que o respectivo polinómio interpolador, de grau 8, é

$$\begin{aligned} p(x,y) \quad &= 1/13824\,(13824\,x - 13248\,y + 5228\,x\,y - 3174\,x^2\,y - 828\,x^3\,y - \\ &\quad 138\,x^4\,y - 9504\,y^2 + 1658\,x\,y^2 - 2277\,x^2\,y^2 - 594\,x^3\,y^2 - \\ &\quad -99\,x^4\,y^2 - 4032\,y^3 + 1324\,x\,y^3 - 966\,x^2\,y^3 - 252\,x^3\,y^3 - \\ &\quad 42\,x^4\,y^3 - 864\,y^4 + 430\,x\,y^4 - 207x^2\,y^4 - 54\,x^3\,y^4 - 9\,x^4\,y^4)\ . \end{aligned}$$

No domínio considerado, a superfície correspondente a $p(x, y)$ acompanha bastante melhor a superfície definida por $f(x, y)$, conforme se pode obervar na Figura 4.12.

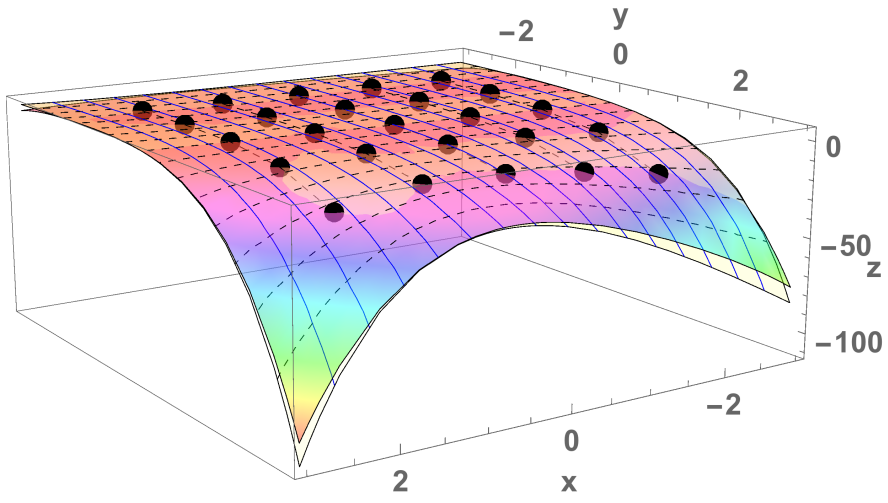


Figura 4.12: Polinómio interpolador para uma malha quadrada de $N = 25$ pontos, em $[-2, 2] \times [-2, 2]$.

4.3 Método dos mínimos quadrados

A interpolação polinomial introduzida no parágrafo 4.1, pág. 183, tem o inconveniente de ser extremamente sensível a erros nos dados iniciais. Na realidade, a matriz de Vandermonde, referida na página 185, pode ser extremamente mal condicionada, e tanto pior quanto maior for o grau de interpolação n, podendo o seu número de condição, como se viu, atingir valores muito elevados, mesmo para valores de n moderados.

Tal significa que um pequeno desvio num dos valores de f pode resultar num polinómio que não é sequer interpolador ou que apresenta oscilações de grande amplitude. Esta propriedade é altamente indesejável nas aplicações, já que é

frequente trabalharmos com tabelas de valores que são afectados de erros (resultantes de medições ou de arredondamentos). Deste ponto de vista, as fórmulas baricêntricas anteriormente referidas apresentam geralmente propriedades de estabilidade numérica superiores à fórmula interpoladora de Newton.

Por outro lado, como foi observado quando estudámos o erro de interpolação, este erro pode ampliar-se quando se aumenta o grau do polinómio, como se constatou no exemplo de Runge tratado no parágrafo 4.1.9, pág. 207. Tudo isto nos alerta para o facto de que a interpolação polinomial pode não ser uma boa forma de aproximar funções, sobretudo quando o número de dados é elevado, conforme se ilustrou no Exemplo 4.2, pág. 190.

Nesta secção vamos estudar um método alternativo para aproximar funções num conjunto discreto de dados, designado por *método dos mínimos quadrados*. Tal como no caso da interpolação, os dados são constituídos por um determinado suporte. No entanto, aqui disporemos de informação redundante, isto é, um número maior de equações relativamente ao número de incógnitas a determinar.

Entre as vantagens deste método contam-se:

1. Permitir uma grande variedade de funções ajustadoras, sem que a forma da função dependa do número de dados.

2. Ser menos sensível aos erros dos dados (em comparação com a interpolação).

3. Aumentando o número de dados, geralmente a qualidade da aproximação tende a aumentar.

4. A soma (mínima) dos quadrados dos *desvios* (entendendo-se por desvios as diferenças entre os valores dados e os valores previstos), sendo o critério para a escolha da função ajustadora, constitui um índice para avaliar a qualidade da aproximação

A seguir descreve-se o método dos mínimos quadrados, com ajustamentos lineares, restrito ao caso discreto. No parágrafo 4.3.4, pág. 231, far-se-á uma breve referência ao caso em que as funções ajustadoras são não lineares nos parâmetros a determinar.

4.3.1 Ajustamentos lineares no caso discreto

O caso discreto caracteriza-se pela forma como é dada a função a aproximar, ou seja, através de uma tabela de pontos (tal como no caso da interpolação polinomial).

Fixado $n \geq 1$, sejam $f_i = f(x_i)$ valores de uma função f nos pontos x_i ($i = 0, 1, \ldots, n$). O objectivo é construir uma determinada função g, dita *função*

ajustadora, definida num intervalo que contém os pontos dados e que constitui, num certo sentido a especificar adiante, a melhor aproximação de f entre a classe de funções que escolhemos como funções aproximantes.

A função ajustadora depende de um certo número de parâmetros, que representaremos genericamente por a_0,a_1,...,a_m.

No caso dos *ajustamentos lineares*, de que trataremos em primeiro lugar (entenda-se linearidade no que respeita aos parâmetros), a função ajustadora pertence a um espaço linear de funções de dimensão $m+1$, podendo ser escrita na forma

$$g(x) = \sum_{i=0}^{m} a_i \phi_i(x), \tag{4.56}$$

onde ϕ são funções dadas, chamadas as *funções de base*. As funções de base devem estar definidas em todos os pontos x_i e devem, além disso, ser linearmente independentes, no seguinte sentido: se fizermos corresponder a cada função ϕ_i um vector $\bar{\phi}_i$ tal que $\bar{\phi}_i = (\phi_i(x_0), \ldots, \phi_i(x_n))$, os vectores $\bar{\phi}_i$ são linearmente independentes em $\mathbb{R}^{n+1}$, para $i = 0, 1, \ldots, m$.

Nas aplicações as funções de base são escolhidas levando em atenção certas propriedades da função a aproximar.

Exemplo 4.8. *Se a função a aproximar for linear, as funções de base poderão ser $\phi_0(x) = 1$, $\phi_1(x) = x$, de tal modo que o espaço linear onde se procura a função ajustadora é o espaço das funções da forma*

$$g(x) = a_0 + a_1 x,$$

ou seja, o dos polinómios de grau não superior a 1.

Num contexto mais geral, se quisermos usar como função ajustadora um polinómio de grau m, as funções de base a utilizar poderão ser os monómios

$$\phi_i(x) = x^i, \qquad i = 0 : m \ .$$

Note-se que esta base de funções polinomiais, denominada usualmente como *base canónica*, é constituida por elementos linearmente independentes no sentido acima mencionado, quaisquer que sejam os $(n+1)$ pontos distintos x_i, com $i = 0 : n$, e $n \geq m$, já que os vectores $\bar{\phi}_i$ têm a forma

$$\bar{\phi}_i = (x_0^i, x_1^i, ..., x_n^i), \qquad i = 0 : m,$$

os quais formam um conjunto linearmente independente.

Exemplo 4.9. *No caso da aproximação de funções periódicas é comum usarem-se bases de funções trigonométricas, como por exemplo*

$$\phi_0(x) = 1, \quad \phi_i(x) = \cos(ix), \qquad i = 0 : m \ .$$

Com funções deste tipo, o sistema poderá ser ou não linearmente independente, consoante a escolha dos pontos x_j, e o número de funções de base. Se tivermos, por exemplo, $x_j = j\pi/4$, para $j = 0:4$, os vectores $\bar{\phi}_i$ neste caso têm a forma

$$\begin{array}{l} \bar{\phi}_0 = (1,1,1,1,1) \\ \bar{\phi}_i = (1, \cos(i*\pi/4), \cos(2i*\pi/4), \cos(3i*\pi/4), \cos(i*\pi)), \qquad i = 0:m, \end{array}$$

os quais são linearmente independentes, para $m \leq 4$.

4.3.2 O critério de mínimos quadrados

Uma vez escolhidas as funções de base ϕ_i, determinar a função ajustadora corresponde a determinar os coeficientes a_i da fórmula (4.56). Estes coeficientes são obtidos com base no *critério dos mínimos quadrados*, ou seja, de modo a minimizar a soma

$$Q(a_0, a_1, ..., a_m) = \sum_{i=0}^{n} (f(x_i) - g(x_i))^2. \tag{4.57}$$

Visto que Q representa uma função de $m+1$ variáveis, a solução deste problema de minimização obtém-se resolvendo o sistema

$$\left\{ \begin{array}{ll} \dfrac{\partial Q}{\partial a_0} & = 0 \\ \dfrac{\partial Q}{\partial a_1} & = 0 \\ & \vdots \\ \dfrac{\partial Q}{\partial a_m} & = 0 \ . \end{array} \right. \tag{4.58}$$

dito *sistema normal*, ou sistema de *equações normais*.

Para construir o sistema normal e discutir as suas propriedades é conveniente, antes de mais, introduzir um produto interno. O produto interno usual de vectores de $\mathbb{R}^{n+1}$ é adequado aos fins em vista. Em particular, o produto interno de duas funções u, v, definidas nos pontos x_i da tabela de valores considerada, é dado por

$$\langle u, v \rangle = \sum_{i=0}^{n} u(x_i) v(x_i) \ .$$

Usando a notação anterior, a função Q em (4.57) pode ser reescrita como o produto interno

$$Q(a_0, a_1, ..., a_m) = \langle f - g, f - g \rangle \ . \tag{4.59}$$

Por outro lado, usando as propriedades do produto interno real, as derivadas parciais de Q podem ser representadas do seguinte modo:

$$\frac{\partial Q}{\partial a_i} = \frac{\partial \langle f-g, f-g\rangle}{\partial a_i} = 2\,\langle \frac{\partial (f-g)}{\partial a_i}, f-g\rangle \ . \tag{4.60}$$

Utilizando a expressão (4.56) para g, de (4.60) obtém-se

$$\frac{\partial Q}{\partial a_i} = -2\,\langle \frac{\partial \left(\sum_{j=0}^{m} a_j\phi_j - f_j\right)}{\partial a_i}, f-g\rangle = -2\,\langle \phi_i, f-g\rangle \ .$$

Sendo assim, cada uma das equações do sistema (4.58) pode ser escrita na forma

$$\langle \phi_i, f-g\rangle = 0, \qquad i = 0:m \tag{4.61}$$

ou seja,

$$\langle \phi_i, g\rangle = \langle \phi_i, f\rangle, \qquad i = 0:m \ .$$

Usando mais uma vez a representação (4.56) e a propriedade distributiva do produto interno, obtém-se finalmente

$$\sum_{j=0}^{m} a_j\,\langle \phi_i, \phi_j\rangle = \langle \phi_i, f\rangle, \qquad i = 0:m \tag{4.62}$$

que constitui a forma compacta do chamado *sistema normal*.

A designação *sistema normal* resulta da expressão (4.61), a qual exprime que a melhor aproximação de mínimos quadrados é obtida quando o vector $f-g$ (ou $g-f$) é ortogonal a cada um dos elementos da base ϕ_0, ϕ_1, ..., ϕ_m, ou seja, ao subespaço G de $\mathbb{R}^{n+1}$ gerado por essa base (ver Figura 4.13).

Concluímos assim que o sistema normal é um sistema linear de $m+1$ equações lineares que pode ser escrito na forma

$$\begin{bmatrix} \langle\phi_0,\phi_0\rangle & \langle\phi_0,\phi_1\rangle & \dots & \langle\phi_0,\phi_m\rangle \\ \langle\phi_1,\phi_0\rangle & \langle\phi_1,\phi_1\rangle & \dots & \langle\phi_1,\phi_m\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle\phi_m,\phi_0\rangle & \langle\phi_m,\phi_1\rangle & \dots & \langle\phi_m,\phi_m\rangle \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \langle\phi_0, f\rangle \\ \langle\phi_1, f\rangle \\ \vdots \\ \langle\phi_m, f\rangle \end{bmatrix}. \tag{4.63}$$

A matriz S do sistema normal é simétrica, dado que

$$S_{ij} = \langle \phi_i, \phi_j\rangle = \langle \phi_j, \phi_i\rangle = S_{ji}, \quad \forall i,j \in \{0,...,m\},$$

o que facilita a sua construção, uma vez que basta calcular as entradas da diagonal principal e as que se encontram acima ou abaixo desta.

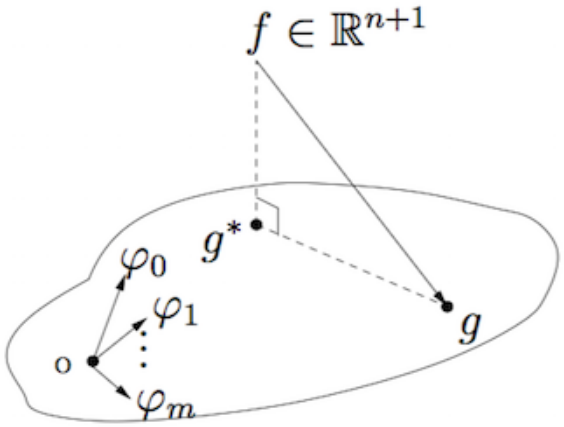


Figura 4.13: O ponto g^* assinala a melhor aproximação de mínimos quadrados.

4.3.3 Unicidade da melhor aproximação de mínimos quadrados

Note-se que as condições

$$\begin{aligned} g(x_0) &= f_0 \\ g(x_1) &= f_1 \\ &\vdots \\ g(x_n) &= f_n \end{aligned}$$

são equivalentes a um sistema $A\,x = f$, sobredeterminado, nas incógnitas $a_0, a_1, \ldots, a_m$, com $f = (f_0, f_1, \ldots, f_n)$ e

$$A = \begin{bmatrix} | & | & & | \\ \phi_0 & \phi_1 & \cdots & \phi_n \\ | & | & & | \end{bmatrix} .$$

Na matriz anterior cada coluna contém as entradas do vector ϕ_i. É fácil concluir que a matriz do sistema de equações normais (4.62) satisfaz a igualdade

$$S = A^T\,A\ .$$

Como por hipótese as colunas de A são linearmente independentes, então para qualquer vector $x \neq 0$, o vector $y = A\,x \neq 0$. Por conseguinte $y^T\,y = ||y||_2^2 = x^T\,A^T\,A\,x = x^T\,Sx > 0$, uma vez que $y \neq 0$.

Conclui-se portanto que a matriz S é *definida positiva* (ver pág. 163) e, consequentemente, o sistema normal $S\,x = A^T\,f$ possui solução única. Assim, a melhor aproximação de mínimos quadrados é *única*.

Em geral, o sistema (4.63) é resolvido numericamente usando, por exemplo, um dos métodos estudados no Capítulo 3. Uma das escolhas mais frequentes é o método de Cholesky, referido na secção 3.2.7, pág. 122, já que este método é aplicável a sistemas de matriz simétrica definida positiva.

Exemplo 4.10. *Se uma determinada grandeza for medida n vezes, erros de observação e/ou de instrumento levam-nos a considerar não o valor exacto dessa grandeza, seja y, mas aproximações (ou "observações") de y,*

$$y_1, y_2, \ldots, y_n \ .$$

Vamos mostrar que a média aritmética das observações é a melhor aproximação de mínimos quadrados da tabela

$$\left\{ \begin{array}{ccccc} 1 & 2 & 3 & \cdots & n \\ y_1 & y_2 & y_3 & \cdots & y_n \end{array} \right\},$$

por funções aproximantes constantes, isto é, do tipo

$$g(x) = c, \qquad c \in \mathbb{R} \ .$$

Com efeito, as "equações de observação",

$$\begin{array}{rl} g(1) & = y_1 \\ g(2) & = y_2 \\ & \vdots \\ g(n) & = y_n, \end{array}$$

traduzem-se no sistema linear incompatível $A\,c = y$, onde

$$A\,c = y \ \Leftrightarrow \ \left[\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array} \right] c = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right].$$

O sistema de equações normais $A^T\,Ac = A^T\,y$, possui solução imediata,

$$A^T\,Ac = A^T\,y \iff n\,c = y_1 + y_2 + \ldots + y_n \iff c = \frac{\sum_{i=1}^{n} y_i}{n} \ .$$

Claro que poderemos chegar à mesma conclusão, considerando o vector de observações $y = (y_1, y_2, \ldots, y_n)$, o vector de base $\bar{\phi}_0 = (1, 1, \ldots, 1)^T$, bem como

o vector de ajustamento $g = c\,\bar{\phi}_0 = (c, c, \ldots, c)^T$. Pretende-se determinar a constante c que minimiza

$$Q(c) = \sum_{i=1}^{n} (g_i - y_i)^2 = \sum_{i=1}^{n} (c - y_i)^2.$$

É condição necessária para que $Q(c)$ possua extremo que $Q'(c) = 0$, isto é,

$$2 \sum_{i=1}^{n} (c - y_i) = 0 \iff \sum_{i=1}^{n} (c - y_i) = 0 \iff n\,c = \sum_{i=1}^{n} y_i \ .$$

Note-se que o mínimo é atingido porquanto $Q''(c) = n > 0$, e este mínimo é único $\forall c \in \mathbb{R}$. Ou seja, a melhor aproximação de mínimos quadrados do suporte dado é a função constante

$$y(x) = c = \frac{\sum_{i=1}^{n} y_i}{n},$$

a qual é igual ao valor da média aritmética das observações. ⧫

O exemplo a seguir ilustra a aplicação do método dos mínimos quadrados discreto escolhendo funções aproximantes do tipo racional.

Exemplo 4.11. *Consideremos a seguinte tabela de valores de uma função f,*

x_i	1	2	3	4
f_i	7	4.5	3	2

Pretende-se aproximar a função f através de uma função ajustadora da forma

$$g(x) = a_0 + \frac{a_1}{x} \ .$$

Trata-se portanto de um ajustamento linear nos parâmetros a_0 e a_1 com duas funções de base,

$$\phi_0(x) = 1, \quad \phi_1(x) = 1/x \ .$$

Para resolver o problema, os valores de a_0 e a_1 podem ser obtidos através do sistema normal,

$$\begin{bmatrix} \langle \phi_0, \phi_0 \rangle & \langle \phi_0, \phi_1 \rangle \\ \langle \phi_1, \phi_0 \rangle & \langle \phi_1, \phi_1 \rangle \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \langle \phi_0, f \rangle \\ \langle \phi_1, f \rangle \end{bmatrix}. \tag{4.64}$$

Calculemos os produtos internos que entram na formação do sistema normal:

$$\langle \phi_0, \phi_0 \rangle = \sum_{i=0}^{3} \phi_0(x_i)^2 = 1+1+1+1 = 4$$

$$\langle \phi_0, \phi_1 \rangle = \sum_{i=0}^{3} \phi_0(x_i)\phi_1(x_i) = \frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} = \frac{25}{12}$$

$$\langle \phi_1, \phi_1 \rangle = \sum_{i=0}^{3} \phi_1(x_i)^2 = \frac{1}{1}+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2} = \frac{205}{144}$$

$$\langle \phi_0, f \rangle = \sum_{i=0}^{3} \phi_0(x_i) f(x_i) = f(1)+f(2)+f(3)+f(4) = 16.5$$

$$\langle \phi_1, f \rangle = \sum_{i=0}^{3} \phi_1(x_i) f(x_i) = f(1)+\frac{f(2)}{2}+\frac{f(3)}{3}+\frac{f(4)}{4} = 10.75 \ .$$

Substituindo estes valores no sistema (4.64), obtém-se

$$\begin{bmatrix} 4 & 25/12 \\ 25/12 & 205/144 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 16.5 \\ 10.75 \end{bmatrix}.$$

A solução do sistema anterior é $a_0 = 0.8077$ e $a_1 = 6.369$, pelo que a função ajustadora procurada é

$$g(x) = 0.8077 + \frac{6.369}{x} \ .$$

Vamos calcular o mínimo

$$\min_{(a_0,a_1)\in\mathbb{R}^2} Q(a_0,a_1) = \min_{(a_0,a_1)\in\mathbb{R}^2} \sum_{i=0}^{3} (f(x_i)-g(x_i))^2 = \sum_{i=0}^{3} (f(x_i)-a_0-a_1/x_i)^2.$$

De acordo com os cálculos já efectuados, este mínimo é atingido quando $a_0 = 0.8077$ e $a_1 = 6.369$, pelo que basta calcular $Q(0.8077, 6.369)$. O resumo dos cálculos é apresentado na tabela a seguir.

x_i	f_i	$g(x_i)$	$d_i^2 = (f_i - g(x_i))^2$
1	7	7.177	0.031
2	4.5	3.992	0.258
3	3	2.931	0.005
4	2	2.400	0.160

O valor procurado é a soma dos valores da última coluna da tabela,

$$Q(0.8077, 6.369) = 0.454 \ .$$

Note que esta coluna contém os quadrados dos *desvios* $d_i = f_i - g_i$.

Conforme resulta da definição do método, é válida a desigualdade

$$Q(a_0, a_1) \geq 0.454, \qquad \forall\, a_0, a_1 \ \in \mathbb{R} \ .$$

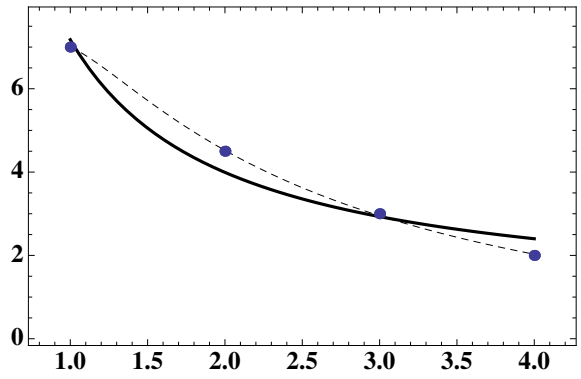


Figura 4.14: Os pontos representam os valores de f e a linha contínua o gráfico da melhor aproximação de mínimos quadrados do tipo $a_0 + a_1/x$. A tracejado a melhor aproximação por funções do tipo $g(x) = a_0 + a_1\, x + a_2/x^2$ (ver Exemplo 4.11).

A Figura 4.14 ilustra a localização da melhor aproximação por funções do tipo referido. Na mesma figura encontra também traçado o gráfico da melhor aproximação de mínimos quadrados que pode ser obtida mediante funções aproximações racionais do tipo

$$g(x) = a_0 + \frac{a_1}{x} + \frac{a_2}{x^2}\ .$$

Pode verificar que a melhor aproximação de mínimos quadrados é aproximadamente,

$$g(x) \simeq -1.301 + \frac{14.99}{x} - \frac{6.690}{x^2}\ .$$

♦

4.3.4 O caso não linear

No parágrafo anterior consideramos apenas funções aproximantes lineares nos parâmetros. Caso o modelo de funções aproximantes seja não linear, somos levados a resolver um sistema *não linear* a fim de determinarmos a melhor aproximação de mínimos quadrados de um dado suporte. Para o efeito, serão úteis os métodos estudados no Capítulo 3, nomeadamente o método de Newton (ver secção 3.7.2, pág. 175).

O Exemplo 4.12 a seguir ilustra um caso em que se compara a abordagem de mínimos quadrados por funções aproximantes lineares, com aproximantes não lineares nos respectivos parâmetros.

t_i	0	1	3	5	7	9
y_i	1.0	0.891	0.708	0.562	0.447	0.355

Tabela 4.7: Ver Exemplo 4.12.

Exemplo 4.12. *Pretende-se optar pela melhor aproximação da Tabela 4.7 por funções aproximantes do tipo polinomial parabólico, ou por funções do tipo exponencial (não lineares nos parâmetros), nomeadamente por funções*

$$g(t) = a_0 + a_1\, t + a_2\, t^2, \tag{4.65}$$

ou

$$h(t) = a\, e^{b\, t}. \tag{4.66}$$

Será adoptada como mais satisfatória a melhor aproximação de mínimos quadrados da tabela para a qual seja menor a soma dos respectivos quadrados dos desvios (ou resíduos).

Para as funções aproximantes do tipo parabólico (função g), pode estabelecer-se o seguinte sistema de equações normais,

$$\begin{bmatrix} 6 & 25 & 165 \\ 25 & 165 & 1225 \\ 165 & 1225 & 9669 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 3.963 \\ 12.149 \\ 71.971 \end{bmatrix},$$

cuja solução é $(0.996954, -0.107378, 0.00403475)^T$. Assim, a melhor aproximação polinomial quadrática da tabela é a função

$$g(t) = 0.996954 - 0.107378\, t + 0.00403475\, t^2.$$

Na Figura 4.15 é mostrado o gráfico de $g(t)$ bem como a respectiva soma dos quadrados dos desvios, $\sum_{i=0}^{5}(g(t_i) - y_i)^2 = 0.0000485629$.

Antes de passarmos ao cálculo da melhor aproximação por funções do tipo (4.66), note-se que se fizermos

$$\ln(h(t)) = \ln(a) + b\, t,$$

poderemos lidar com funções aproximantes lineares do tipo

$$Y(t) = a_0 + a_1\, t, \quad \text{com} \quad a_0 = \ln(a) \quad \text{e} \quad a_1 = b\ . \tag{4.67}$$

Faz por isso sentido, começar por calcular a melhor aproximação linear por funções do tipo (4.67), da Tabela 4.8. e usar os parâmetros que resultam dessa aproximação como estimativa inicial dos parâmetros a determinar para as funções aproximantes do tipo (4.66).

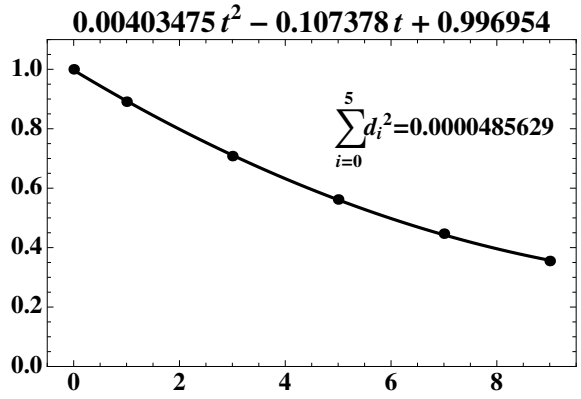


Figura 4.15: Melhor aproximação polinomial quadrática da Tabela 4.7.

t_i	0	1	3	5	7	9
$\ln(y_i)$	0	-0.115411	-0.345311	-0.576253	-0.805197	-1.03564

Tabela 4.8: Valores de $\ln(y_i)$, a partir da Tabela 4.7.

O sistema normal a resolver levando em consideração os dados da Tabela 4.8, é

$$\begin{bmatrix} 6 & 25 \\ 25 & 165 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} -2.87781 \\ -18.9897 \end{bmatrix},$$

cuja solução é $(-0.000261498, -0.11505)^T$. Assim, a melhor aproximação polinomial linear dessa tabela é a função

$$Y(t) = -0.000261498 - 0.11505\, t\ .$$

A respectiva soma dos quadrados dos desvios é $\sum_{i=0}^{5}(Y(t_i) - \ln(y_i))^2 = 8.16301 \times 10^{-7}$.

Como se disse anteriormente, os valores dos parâmetros $a_0 = \ln(a)$ e $a_1 = b$, servem-nos agora como *aproximação inicial* dos parâmetros a e b, tendo em vista o cálculo da melhor aproximação não linear da tabela original por funções do tipo h. Assim,

$$\begin{aligned} a &\simeq e^{a_0} = 0.999739 \\ b &\simeq -0.11505\ . \end{aligned}$$

Passemos agora ao cálculo da aproximação não linear de mínimos quadrados. Para minimizarmos

$$Q(a,b) = \sum_{i=0}^{5}(h(t_i) - y_i)^2 = \sum_{i=0}^{5}(a\, e^{b\, t_i} - y_i)^2,$$

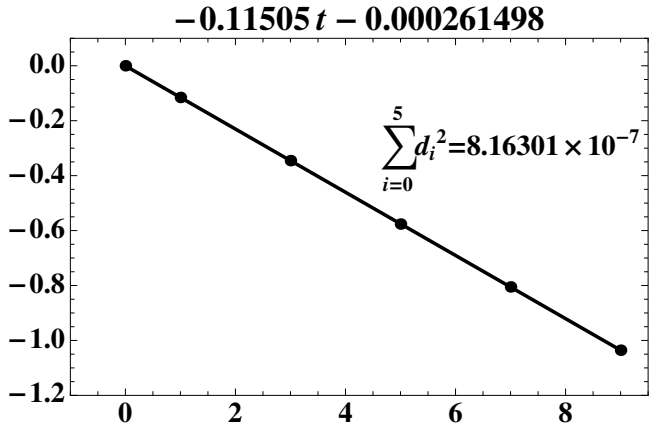


Figura 4.16: Melhor aproximação linear da tabela 4.8 .

tem-se,

$$\begin{aligned}\frac{\partial}{\partial a}Q(a,b) &= \sum_{i=0}^{5}(a\,e^{b\,t_i}-y_i)\,e^{b\,t_i} &&= 0\\ \frac{\partial}{\partial b}Q(a,b) &= \sum_{i=0}^{5}(a\,e^{b\,t_i}-y_i)\,a\,t_i\,e^{b\,t_i} &&= 0\ .\end{aligned}$$

Por conseguinte, o sistema não linear a resolver é da forma,

$$\left\{\begin{array}{ll}(\sum_{i=0}^{5}\,e^{2\,b\,t_i})\,a-\sum_{i=0}^{5}y_i e^{\,b\,t_i} & =0\\ (\sum_{i=0}^{5}t_i\,e^{2\,b\,t_i})\,a^2-(\sum_{i=0}^{5}y_i\,t_i e^{\,b\,t_i})\,a & =0\ .\end{array}\right.$$

O leitor pode verificar que fazendo $X^{(0)}=(0.999739,-0.11505)$, a primeira iterada do método de Newon aplicado ao sistema anterior produz o resultado

$$X^{(1)}=(0.999841,-0.115083),$$

a qual coincide com a iterada $X^{(2)}$ (para a precisão utilizada nos cálculos). Assim, a melhor aproximação de mínimos quadrados da tabela inicial, por funções do tipo (4.66) tem aproximadamente a forma

$$h(t)\simeq 0.999841\,e^{-0.115083\,t}.$$

O gráfico de $h(t)$ é mostrado na Figura 4.17. A respectiva soma dos quadrados dos desvios é $\sum_{i=0}^{5}(h(t_i)-y_i))^2=2.66547\times10^{-7}$. Comparando com a soma dos quadrados dos desvios anteriormente calculada para o ajuste polinomial parabólico, concluimos que a aproximação não linear calculada é mais precisa (embora exija um esforço computacional muito maior).

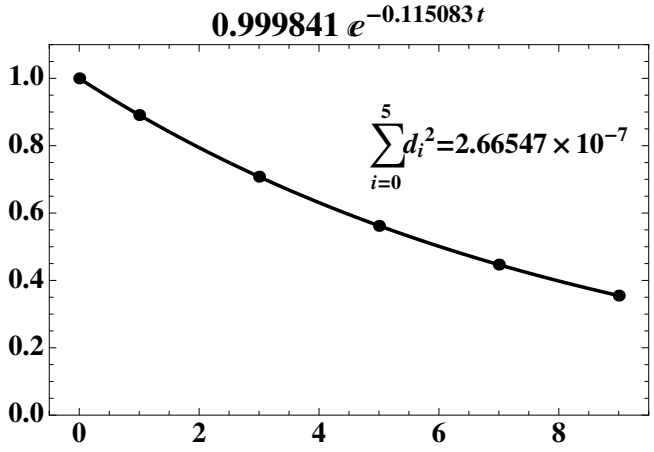


Figura 4.17: Melhor aproximação polinomial quadrática da tabela 4.7 por funções do tipo h .

4.4 Exercícios resolvidos

Exercício 4.2. *Considere a seguinte tabela de valores da função* $f(x) = \log_{10}(x)$*:*

x_i	2.0	2.5	3.0
$\log_{10}(x_i)$	0.30103	0.39794	0.47712

(a) Calcule uma aproximação de $\log_{10}(2.4)$*, usando a fórmula interpoladora de Newton e todos os pontos da tabela.*

(b) Estime o erro de interpolação em $x = 2.4$ *e compare-o com o erro exacto.*

(c) Determine um majorante do erro absoluto que se comete em $[2.0, 3.0]$ *ao aproximar a função* f *pelo polinómio que obteve na alínea (a).*

(d) Substitua a tabela dada por aquela que se obtém considerando os nós

$$x_0 = 2 + \frac{1}{2}\left(1 - \frac{\sqrt{3}}{2}\right), \quad x_1 = 2.5, \quad x_2 = 2 + \frac{1}{2}\left(1 + \frac{\sqrt{3}}{2}\right).$$

Obtenha o gráfico da função erro de interpolação $e_2(x) = f(x) - P_2(x)$*, onde* $P_2(x)$ *designa o polinómio interpolador do suporte de interpolação que tenha como nós* x_0*,* x_1 *e* x_2 *(estes nós resultam de uma translação dos zeros do polinómio de Chebyshev de grau 3 (ver pág. 211)).*
O erro de interpolação global de P_2 *é ou não é menor do que aquele que calculou na alínea (c)?*

(a) A partir do suporte de interpolação dado, construa-se a seguinte tabela de

diferenças divididas:

x_i	f_i	$f[\cdot\,\cdot]$	$f[\cdot\,\cdot\,\cdot]$
2.0	0.30103		
		0.19382	
2.5	0.39794		-0.03546
		0.15836	
3.0	0.47712		

O polinómio interpolador de Newton tem a forma,

$$\begin{aligned} p_2(x) &= f[2.0] + f[2.0, 2.5]\,(x - 2.0) + f[2.0, 2.5, 3.0]\,(x - 2.0)\,(x - 2.5) \\ &= 0.30103 + 0.19382\,(x - 2.0) - 0.03546\,(x - 2.0)\,(x - 2.5)\ . \end{aligned}$$

Fazendo $x = 2.4$, obtém-se

$$p_2(2.4) \simeq 0.379976\ .$$

O valor anterior aproxima $\log_{10}(2.4) = 0.380211$ (6 algarismos significativos) com um erro de interpolação (exacto dentro da precisão usada nos cálculos) de

$$e_2(2.4) = f(2 - 4) - p_2(2.4) = 0.000235\ .$$

(b) A função $f(x) = \log_{10}(x)$, no intervalo $I = [2.0, 3.0]$, é suficientemente regular, pelo que é aplicável a fórmula teórica de erro (4.31), pág. 204, para interpolação parabólica, isto é, para $n = 2$,

$$e_2(x) = f(x) - p_2(x) = \frac{f^{(3)}(\xi)}{3!}\,(x-2.0)\,(x-2.5)\,(x-3.0), \quad \xi = \xi(x) \in (2.0, 3.0)\ . \tag{4.68}$$

Fixado $x = 2.4$, uma majoração do erro local de interpolação pode escrever-se como

$$e = |e_2(2.4)| \ \leq M \times |(2.4 - 2.0)\,(2.4 - 2.5)\,(2.4 - 3.0)|, \tag{4.69}$$

onde

$$M = \frac{1}{3!}\,max_{x\in[2.0,3.0]}\,|f^{(3)}(x)|\ .$$

Como,

$$\begin{aligned} &f(x) = \log_{10}(x) = \frac{\ln(x)}{\ln(10)} = c\,\ln(x), \quad \text{com} \quad c = 1/\ln(10), \\ &f'(x) = \frac{c}{x}, \quad f^{(2)}(x) = -\frac{c}{x^2}, \quad f^{(3)}(x) = \frac{2\,c}{x^3}, \end{aligned}$$

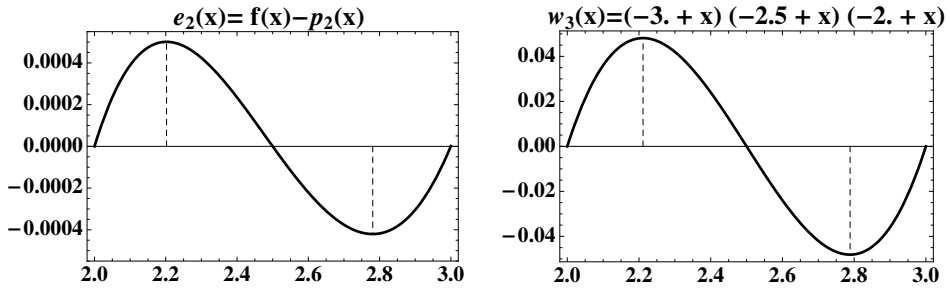


Figura 4.18: Os gráficos de $e_2(x)$ e $w_2(x)$.

no intervalo I considerado, a função $f^{(3)}(x)$ é positiva e estritamente decrescente, pelo que o seu máximo ocorre no extremo esquerdo do intervalo. Assim,

$$M = \frac{1}{3!} f^{(3)}(2.0) = \frac{1}{3!}\,\frac{2\,c}{2^3} = \frac{c}{24} \simeq 1.80956 \times 10^{-2}.$$

Substituindo em (4.69), obtém-se

$$e \leq 1.80956 \times 10^{-2} \times 0.4 \times 0.1 \times 0.6 \simeq 0.000434\ .$$

A majoração de erro assim obtida é aproximadamente duas vezes superior ao erro de interpolação efectivamente cometido.

Conforme a expressão (4.68) sugere, o erro de interpolação depende de f, do ponto $x \in I$ considerado, e dos nós de interpolação. Rescrevendo (4.68) na forma

$$e_2(x) = \frac{f^{(3)}(\xi)}{3!}\, w_3(x), \quad \xi = \xi(x) \in (2.0, 3.0),$$

evidenciamos que o factor polinomial $w_3(x)$[8] da expressão do erro depende da localização de x relativamente aos nós x_0, x_1, x_2.

Designando por E o erro máximo de interpolação cometido em todo o intervalo, ou seja o erro máximo global, tem-se

$$E = max_{x\in I}|f(x) - p_2(x)| \leq M \times max_{x\in I}\,|w_3(x)|\ . \tag{4.70}$$

Na Figura 4.18 são comparados os gráficos da função $e_2(x) = f(x) - p_2(x)$, e do polinómio nodal $w_3(x)$. Note que os pontos de extremos de $w_3(x)$ aproximam os pontos de extremos de $e_2(x)$. De facto, $e_2(x)$ tem valores extremos próximo de $x = 2.20$ e $x = 2.78$, enquanto que os extremos de $w_3(x)$ ocorrem próximo dos pontos $x = 2.21$ e $x = 2.79$.

[8]Relembre-se que o polinómio $w_{n+1}(x)$ é por vezes designado como *polinómio nodal* por estar associado aos nós de interpolação.

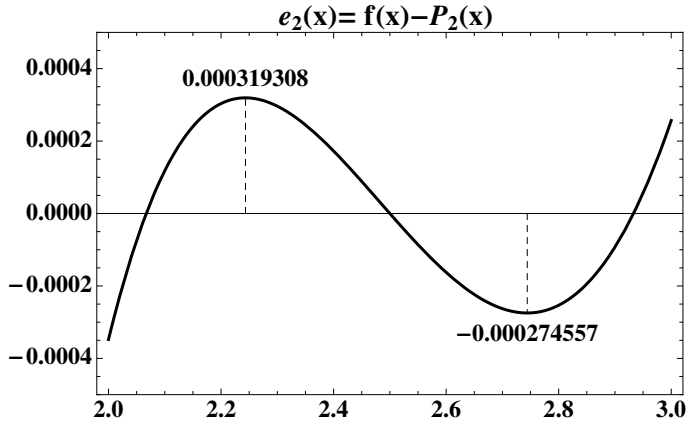


Figura 4.19: O gráfico de $e_2(x)$ para um suporte de Chebyshev.

Neste caso, podemos determinar expressões exactas para a localização dos extremos de $w_3(x)$, visto que a derivada $w_3'(x)$ é um polinómio do segundo grau. Com efeito,

$$w_3'(x) = (x-2.5)\,(x-3.0) + (x-2.0)\,(x-3.0) + (x-2.0)\,(x-2.5)\ .$$

Ou seja,

$$\begin{array}{l} x^2 - 5.5\,x + 7.5 \\ x^2 - 5.0\,x + 6.0 \\ x^2 - 4.5\,x + 5.0 \\ \hline w_3'(x) = 3\,x^2 - 15\,x + 18.5, \qquad \in \mathcal{P}_2\ . \end{array}$$

Assim, $w_3'(x) = 0 \iff x_{1,2} = \dfrac{15 \pm \sqrt{15^2 - 12 \times 18.5}}{6}$.
Designando os zeros de w_3' por α_1 e α_2, tem-se

$$\begin{aligned} \alpha_1 \simeq 2.21132 \ &\Rightarrow\ w_3(\alpha_1) = \frac{1}{12\sqrt{3}} \simeq 0.0481125 \\ \alpha_2 \simeq 2.78868 \ &\Rightarrow\ w_3(\alpha_2) = -\frac{1}{12\sqrt{3}} \simeq -0.0481125\ . \end{aligned}$$

Note-se que α_1 e α_2 são simétricos relativamente ao nó central $x_1 = 2.5$.

Por conseguinte, a majoração (4.70) de erro de interpolação global passa a ser,

$$\begin{aligned} E \ &\leq M \times w_3(\alpha_1) \\ &\leq 1.80956 \times 10^{-2} \times 0.0481125 \simeq 0.000871, \end{aligned}$$

isto é, o erro máximo global E é aproximadamente o dobro do erro local e calculado na alínea (b).

(d) Constatamos nesta alínea como o erro de interpolação é susceptível de variar em função da escolha feita dos nós de interpolação. Aqui foram adoptados os nós

de Chebyshev, os quais minoram o factor nodal, $w_3(x)$, que faz parte da fórmula teórica de interpolação que referimos anteriormente.

O novo suporte de interpolação (para seis algarismos significativos) é dado na seguinte tabela:

x_i	2.06699	2.5	2.93301
$\log_{10}(x_i)$	0.315338	0.397940	0.467314

O respectivo polinómio interpolador (na base canónica), tem a forma

$$P_2(x) = -0.261248 + 0.351864\,x - 0.0352754\,x^2 \ .$$

Na Figura 4.19 está representada a função erro de interpolação, $e_2(x) = f(x) - P_2(x)$. O erro máximo absoluto ocorre em $x = 2.24344$, e o seu valor é de $E = 0.000319308$, o qual vale cerca de metade do erro máximo de interpolação calculado na alínea anterior. ⧫

4.5 Leituras aconselhadas

J.-P. Berrut, L. N. Trefethen, Barycentric Lagrange interpolation, *SIAM Rev.*, 46(3), 501-517, 2004.

J.-P. Berrut, Fascinante interpolation, *Bull. Soc. Frib. Sc. Nat.*, 83(1/2), 3-20, 1994.

J. P. Boyd, Finding the zeros of a univariate equation: Proxy root finders, Chebyshev interpolation, and the companion matrix, SIAM Rev., 55(2), 375-396, 2013.

J. P. Boyd, *Solving transcendental equations, the Chebyshev Polynomial Proxy and other numerical root finders, perturbation series, and oracles*, SIAM, Philadelphia, 2014.

The Discovery of Ceres, in *Kepler's Discovery*,
http://www.keplersdiscovery.com/Asteroid.html.

N. Crato, O papel dos mínimos quadrados na descoberta dos planetas, *Boletim SPM* 42, 113-124, 2000.

J. F. Epperson, *On the Runge Example*, 1987,
http://www.maa.org/sites/default/files/images/upload_library/22/Ford/Epperson329-341.pdf

A. Gil, J. Segura, and N. Temme, *Numerical Methods for Special Functions*, Ch. 3, SIAM, Philadelphia, 2007, (disponível em:
http://www.siam.org/books/ot99/OT99SampleChapter.pdf).

George M. Phillips, *Interpolation and Approximation by Polynomials*, Canadian Mathematical Society, Springer, New York, 2003.

H. Pina, *Métodos Numéricos*, Escolar Editora, 2010, Cap. 2.

Capítulo 5

Integração numérica

Neste capítulo trataremos do cálculo aproximado de integrais definidos.

Sendo f uma função real, definida e integrável num certo intervalo $[a, b]$, representaremos por $I(f)$ o integral

$$I(f) = \int_a^b f(x)dx \ .$$

Como é sabido, as fórmulas do cálculo integral que permitem obter analiticamente $I(f)$ só se aplicam a classes restritas de funções (aquelas cuja primitiva é conhecida), pelo que é de grande importância prática o desenvolvimento de métodos numéricos que permitam obter valores aproximados do integral.

Alguns métodos dessa natureza são conhecidos desde a Antiguidade. Por exemplo, Arquimedes,[1] desenvolveu técnicas de integração que utilizou para calcular áreas e volumes de sólidos geométricos.

Designaremos por *regras de quadratura* ou *regras de integração numérica*, certas fórmulas que visam obter aproximações do integral $I(f)$.

Fixado um número $n \geq 0$, o primeiro passo para a construção de uma regra de quadratura consiste na selecção de um certo conjunto de pontos x_i, com $i = 0, 1, \ldots, n$, pertencentes ao intervalo $[a, b]$, a que chamaremos os *nós de integração*.[2]

Para o cálculo aproximado de $I(f)$, usaremos a informação dos valores da função integranda nesses nós. Ou seja, tal como fizemos em interpolação, consideramos o *suporte* $\{x_i, f(x_i)\}_{i=0}^{i=n}$. Uma regra de quadratura, que denotaremos por $I_n(f)$,

[1] Arquimedes de Siracusa, c. 287 AC – c. 212 AC, matemático, físico, astrónomo e engenheiro grego.

[2] Podem igualmente construir-se regras de quadratura com nós exteriores ao intervalo $[a, b]$, o que não faremos aqui.

ou $Q_n(f)$ (ou por símbolos relacionados com o nome adoptado para a regra em causa), terá a forma

$$I_n(f) = \sum_{i=0}^{n} A_i f(x_i), \tag{5.1}$$

onde os coeficientes A_i são números (geralmente positivos), a que chamamos os *pesos* da quadratura.

Os pesos de uma regra de quadratura serão determinados de acordo com os nós de integração fixados e a precisão que se pretende alcançar. Estudaremos neste capítulo algumas técnicas elementares para o seu cálculo, a partir do polinómio interpolador do suporte adoptado, ou resolvendo determinados sistemas lineares.

5.0.1 Integração do polinómio interpolador

Uma forma natural de aproximar um integral definido consiste em substituir o integral da função pelo integral do seu polinómio interpolador, utilizando as fórmulas de interpolação estudadas no capítulo anterior. Veremos adiante que os pesos de uma regra de quadratura podem ser mais facilmente calculados resolvendo certos sistemas lineares.

Considere-se

$$I_n(f) = \int_a^b P_n(x)dx, \tag{5.2}$$

onde P_n é o polinómio que interpola f nos nós $x_0, x_1, ..., x_n$. Uma vez que P_n é interpolador no suporte considerado, é de esperar que $I_n(f)$ seja uma aproximação de $I(f)$. A qualidade dessa aproximação depende da proximidade do polinómio interpolador relativamente à função f, no intervalo $[a, b]$.

Podemos recorrer à fórmula de interpolação de Lagrange (ver pág. 187). Sabemos que

$$P_n(x) = \sum_{i=0}^{n} f(x_i)\, L_i(x), \tag{5.3}$$

onde L_i representa o i-ésimo polinómio de Lagrange. Substituindo (5.3) em (5.2), obtém-se

$$I_n(f) = \int_a^b \sum_{i=0}^{n} f(x_i) L_i(x)\, dx\ . \tag{5.4}$$

Aplicando à expressão (5.4) a propriedade de linearidade dos integrais definidos, temos

$$I_n(f) = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)\, dx\ . \tag{5.5}$$

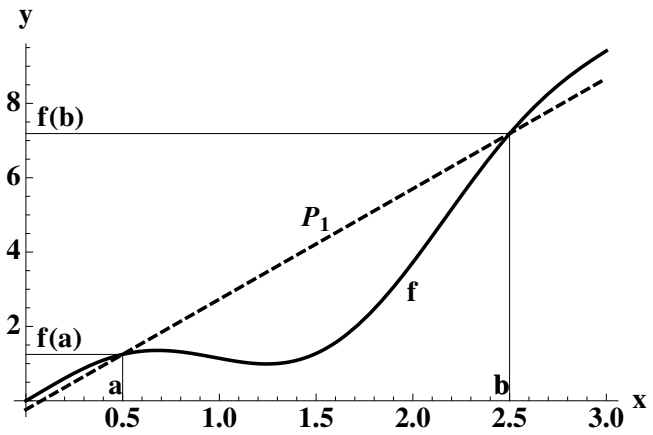


Figura 5.1: Pretende-se calcular o integral de f no intervalo $[0.5, 2.5]$. O valor aproximado do integral, obtido pela regra dos trapézios, é igual à área do trapézio delimitado pelo gráfico de P_1, pelo eixo das abcissas, e pelas rectas $x = 0.5$ e $x = 2.5$ (ver Exemplo 5.1).

Comparando as fórmulas (5.5) e (5.1), deduzimos que os pesos A_i da regra de integração I_n podem ser obtidos calculando os integrais

$$A_i = \int_a^b L_i(x)\, dx, \qquad i = 0, 1, ..., n \ . \tag{5.6}$$

Veremos adiante fórmulas computacionalmente mais económicas para determinar os pesos A_i.

Exemplo 5.1. *Consideremos o caso simples de uma regra de integração com dois nós, $x_0 = a$ e $x_1 = b$,*

$$I_1(f) = A_0 f(a) + A_1 f(b) \ . \tag{5.7}$$

Pretende-se determinar os pesos A_0 e A_1.

Trata-se de aproximar o integral da função f pelo integral do polinómio que interpola f nos nós $x_0 = a$, $x_1 = b$, o qual, como sabemos, é um polinómio de grau não superior a 1,

$$I_1(f) = \int_a^b P_1(x)\, dx \ .$$

Essa aproximação está ilustrada na Figura 5.1. Para calcular os pesos, utilizando a fórmula (5.6), começamos por construir os polinómios de Lagrange. De acordo com a fórmula (4.7), pág. 188, temos

$$L_0(x) = \frac{x-b}{a-b} \qquad \text{e} \qquad L_1(x) = \frac{x-a}{b-a} \ .$$

Aplicando as igualdades (5.6), e calculando analiticamente os integrais dos polinómios de Lagrange, resulta

$$A_0 = \int_a^b \frac{x-b}{a-b}\,dx = \frac{b-a}{2},$$

$$A_1 = \int_a^b \frac{x-a}{b-a}\,dx = \frac{b-a}{2}\ .$$

Finalmente, substituindo em (5.7) os valores dos pesos A_0 e A_1, obtém-se

$$I_1(f) = \frac{b-a}{2}\,f(a) + \frac{b-a}{2}\,f(b) \quad = \frac{h}{2}(f(a)+f(b)), \tag{5.8}$$

para $h = (b-a)$. Esta é uma das fórmulas clássicas da integração numérica, conhecida como *regra dos trapézios*. Na próxima secção, estudá-la-emos mais detalhadamente ⧫

5.1 Regra dos trapézios simples

Embora a fórmula (5.6) seja teoricamente adequada para se calcular os pesos de uma quadratura, ela não é frequentemente a mais eficiente. Existem outras maneiras para determinarmos os pesos, com menos cálculos.

Por exemplo, no caso da regra dos trapézios, poderíamos chegar à fórmula (5.8) simplesmente através do seu significado geométrico. Com efeito, observando a Figura 5.1, facilmente se conclui que o valor de $I_1(f)$ corresponde à área de um trapézio, cuja altura é $h = b-a$, e cujas bases são $f(a)$ e $f(b)$. Daqui poderíamos imediatamente escrever

$$T(f) = (b-a)\frac{f(a)+f(b)}{2} = \frac{h}{2}(f(a)+f(b)), \tag{5.9}$$

expressão idêntica a (5.8), e daí passarmos a designar a regra por $T(f)$, pois $I_1(f) = T(f)$. Foi precisamente o seu significado geométrico que deu o nome à regra dos trapézios.

O passo seguinte no estudo de uma regra de integração consiste na análise do respectivo *erro de quadratura*.

5.1.1 Erro de quadratura

É natural chamar-se *erro de quadratura* ou *erro de integração* à diferença,

$$E_n(f) = I(f) - I_n(f) = \int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx\ .$$

Para a regra dos trapézios, em particular, temos:

$$\begin{aligned} E_T(f) &= I(f) - T(f) = \int_a^b f(x)\,dx - \int_a^b P_1(x)\,dx \\ &= \int_a^b (f(x) - P_1(x))\,dx, \end{aligned} \tag{5.10}$$

ou seja, o erro de integração é igual ao integral do erro de interpolação, quando em vez da função f usamos o polinómio interpolador P_1, no intervalo $[a, b]$. Para calcular esse integral analiticamente podemos recorrer ao erro de interpolação, pág. 202.

Se admitirmos que $f \in C^2([a, b])$, com base em (4.31), pág. 204, sabemos que existe pelo menos um ponto $\xi = \xi(x)$ em (a, b), tal que

$$f(x) - P_1(x) = \frac{f''(\xi(x))}{2}(x - a)(x - b)\ . \tag{5.11}$$

Substituindo (5.11) em (5.10), obtém-se

$$E_T(f) = \int_a^b (f(x) - P_1(x))\,dx = \int_a^b \frac{f''(\xi(x))}{2}(x - a)(x - b)\,dx\ . \tag{5.12}$$

Finalmente, para estimar o integral (5.12), recorre-se a um teorema clássico do cálculo, o chamado *teorema do valor médio para integrais* (ver, por exemplo [29], p. 172).

Segundo o teorema do valor médio para integrais, ao integrar o produto de duas funções u e v num certo intervalo $[a, b]$. Sendo a função u contínua, a função v de *sinal constante* em $[a, b]$, e o produto $u(x)\,v(x)$ integrável, existe pelo menos um ponto η tal que $a \leq \eta \leq b$, para o qual é válida a igualdade,

$$\int_a^b u(x)v(x)dx = u(\eta)\int_a^b v(x)dx\ . \tag{5.13}$$

Para aplicarmos o resultado anterior ao erro da regra dos trapézios, consideremos $u(x) = f''(x)/2$ e $v(x) = (x - a)(x - b)$. A continuidade da função u, resulta de admitirmos que $f \in C^2([a, b])$, enquanto que obviamente $v(x) \leq 0$ em $[a, b]$.

Por conseguinte, a aplicação de (5.13) a (5.12) garante-nos a existência de pelo menos um ponto $\eta \in (a, b)$, tal que

$$E_T(f) = \int_a^b \frac{f''(\xi(x))}{2}(x - a)(x - b)\,dx = \frac{f''(\eta)}{2}\int_a^b (x - a)(x - b)\,dx\ . \tag{5.14}$$

Finalmente, o integral no segundo membro de (5.14) pode ser calculado analiticamente, obtendo-se

$$\begin{aligned} E_T(f) &= -\frac{(b - a)^3}{12} f''(\eta), \qquad \eta \in [a, b] \\ &= -\frac{(b - a)}{12}\,h^2\,f''(\eta)\ . \end{aligned} \tag{5.15}$$

Uma vez que o valor de η é, em geral, desconhecido, na prática considera-se a majoração

$$\begin{aligned} |E_T(f)| &\leq \max_{x\in[a,b]} |f''(x)| \frac{(b-a)^3}{12} \\ &\leq \frac{(b-a)}{12} \max_{x\in[a,b]} |f''(x)|\, h^2 \ . \end{aligned} \tag{5.16}$$

A última desigualdade em (5.16) é válida tanto para a regra dos trapézios simples de que aqui nos ocupamos, como para a chamada regra dos trapézios *composta* (ver pág. 248). A referida majoração de erro diz-nos que uma vez fixado o intervalo $[a, b]$, se o subdividirmos num certo número N de partes, de igual comprimento $h = (b-a)/N$ (na regra dos trapézios simples $N = 1$), o erro de quadratura é da ordem do quadrado de h, isto é $|E_T(f)| = \mathcal{O}(h^2)$.

Exemplo 5.2. *Consideremos o integral*

$$I(\cos) = \int_0^{\pi/6} \cos(x)\, dx \ .$$

Pretende-se obter uma aproximação de $I(\cos)$ *e a respectiva estimativa de erro, mediante aplicação da regra dos trapézios simples.*

Para calcularmos um valor aproximado deste integral pela regra dos trapézios basta aplicar a fórmula (5.9),

$$T(f) = \frac{\cos 0 + \cos(\pi/6)}{2} \frac{\pi}{6} \simeq 0.4885 \ .$$

Um majorante do erro desta aproximação, pode obter-se utilizando a fórmula (5.16),

$$|E_T(f)| \leq \max_{x\in[0,\pi/6]} |\cos(x)| \frac{(\pi/6)^3}{12} = \frac{(\pi/6)^3}{12} \simeq 0.0120 \ . \tag{5.17}$$

Atendendo a que $\int_0^{\pi/6} \cos(x)\, dx = 0.5$, o erro de facto cometido é

$$E_T(f) \simeq 0.5 - 0.4885 = 0.0115 \ .$$

Conclui-se que a estimativa dada por (5.17) é, neste caso, bastante realista. ♦

5.1.2 Regra dos trapézios composta

Já tínhamos referido que a regra de quadratura anterior é conhecida por regra dos trapézios *simples*, pois é aplicada no intervalo $[a, b]$ usando apenas dois nós de quadratura (os extremos do intervalo). O "passo" entre nós consecutivos vale, portanto, $h = b - a$.

Como facilmente se depreende da fórmula (5.16), o erro de integração cresce rapidamente com o comprimento do intervalo, pelo que a aproximação só será aceitável para intervalos de comprimento pequeno. Na prática, usa-se a regra dos trapézios *composta*, que passamos a descrever.

Fixado o número inteiro $N \geq 1$, começamos por definir o conjunto de nós equidistantes x_i no intervalo $[a, b]$,

$$x_i = a + i\,h, \qquad h = \frac{b-a}{N}, \qquad i = 0 : N \ .$$

O espaçamento entre nós consecutivos é dado por h (também chamado passo de integração). Relembre que na regra dos trapézios simples $N = 1$.

Decompomos o integral $I(f)$ numa soma de N parcelas,

$$\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \cdots + \int_{x_{N-1}}^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx \ . \tag{5.18}$$

A cada uma das parcelas da soma (5.18) podemos aplicar a regra dos trapézios simples, isto é,

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{f(x_i) + f(x_{i+1})}{2}\, h \ .$$

Assim, o valor total do integral pode ser aproximado pela soma dos valores dados pela fórmula anterior, obtendo-se

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2}\, h \ . \tag{5.19}$$

Facilmente se verifica que o somatório da fórmula (5.19) também pode ser representado na forma

$$\begin{aligned} T_N(f) &= h\left(\frac{f(a)}{2} + f(x_1) + \cdots + f(x_{N-1}) + \frac{f(b)}{2}\right) = h\left(\frac{f(a)}{2} + \frac{f(b)}{2} + \sum_{i=1}^{N-1} f(x_i)\right) \\ &= \frac{h}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{N-1} f(x_i)\right] \ . \end{aligned} \tag{5.20}$$

A fórmula (5.20) é conhecida como regra dos trapézios *composta*, onde o índice N em T_N representa o número de subintervalos considerados.

5.1.3 Estimativa de erro na regra dos trapézios composta

É de esperar que, para um determinado integral, quando se aumenta o número de subintervalos N a aproximação se torne cada vez melhor, isto é, que o erro absoluto de quadratura decresça. Para verificarmos que assim é, vamos obter uma estimativa do erro da regra dos trapézios composta,

$$E_N^T(f) = I(f) - T_N(f) \ .$$

Começamos por observar que o erro de $T_N(f)$ é a soma dos erros cometidos em cada uma dos subintervalos $[x_i, x_{i+1}]$. Se assumirmos que a função f é pelo menos de classe C^2 em $[a, b]$, ou seja, f e as suas primeiras duas derivadas são contínuas em $[a, b]$, o erro de quadratura pode ser avaliado usando a fórmula (5.16), donde

$$|E_N^T(f)| \leq \max_{x\in[x_i,x_{i+1}]} |f''(x)| \ \frac{h^3}{12}, \qquad 0 \leq i \leq (N-1) \ .$$

Somando os erros de integração em todos os sub-intervalos, obtém-se

$$|E_N^T(f)| \leq \sum_{i=0}^{N-1} \max_{x\in[x_i,x_{i+1}]} |f''(x)| \ \frac{h^3}{12} \ . \tag{5.21}$$

Usando a notação $M = \max_{x\in[a,b]} |f''(x)|$, da fórmula (5.21) pode concluir-se que

$$|E_N^T(f)| \ \leq M\, N\, \frac{h^3}{12} = \frac{(b-a)}{12}\, M\, h^2, \quad \text{para} \quad h = \frac{b-a}{N}, \quad M = \max_{x\in[a,b]} |f''(x)| \ . \tag{5.22}$$

A desigualdade anterior é geralmente aplicada para majorar o erro absoluto da regra dos trapézios composta. Conclui-se que, quando $h \to 0$, (isto é, o número de subintervalos $N \to \infty$), o erro de integração tende para zero, ou seja, o valor obtido pela regra converge para o valor exacto do integral.

A fórmula (5.22) poderá servir também para se deduzir qual o valor de h que se deve utilizar se pretendermos calcular o integral com um erro absoluto inferior a uma dada tolerância ϵ ou, equivalentemente, determinarmos qual o número N de subintervalos que devem ser prefixados para satisfazer essa tolerância de erro, tal como é ilustrado no Exemplo 5.3.

Exemplo 5.3. *Consideremos o integral*

$$I(\cos) = \int_0^{\pi/2} \cos(x)\, dx,$$

(a) Pretende-se aproximar o valor de $I(\cos)$ *usando a regra dos trapézios composta, com* 4 *nós de integração, bem como estimar o erro de quadratura correspondente.*

(b) Em quantas partes deveremos subdividir o intervalo de quadratura, de modo a garantir um erro inferior a $\epsilon = 10^{-6}$*?*

(a) O número de subintervalos a considerar é $N = 3$. Logo, o passo é $h = \pi/6$, e os nós de quadratura são

$$x_0 = a = 0, \quad x_1 = \pi/6, \quad x_2 = \pi/3, \quad \text{e} \quad x_3 = \pi/2 \ .$$

Aplicando a fórmula (5.20), obtém-se

$$T_3(f) = \frac{\pi}{6}\left(\frac{\cos(0)}{2} + \cos(\pi/6) + \cos(\pi/3) + \frac{\cos(\pi/2)}{2}\right) \simeq 0.97705 \ .$$

O erro absoluto da aproximação anterior pode ser estimado através da fórmula (5.22). Comecemos por observar que $M = \max_{x\in[0,\pi/2]} |\cos(x)| = 1$. Assim,

$$|E_T^3(f)| \leq M\frac{\pi}{2}\frac{(\pi/6)^2}{12} \simeq 0.0359 \ .$$

Atendendo a que $\int_0^{\pi/2} \cos(x)\,dx = 1$, temos que o erro de facto cometido é

$$E_T^3(f) \simeq 1 - 0.97705 = 0.0229,$$

pelo que a estimativa obtida é bastante realista. Este exemplo é ilustrado na Figura 5.2.

(b) Recorrendo de novo à fórmula (5.22), temos

$$|E_T^N(f)| \leq \frac{\pi}{2}\frac{h^2}{12} \ .$$

Da inequação

$$\frac{\pi}{2}\frac{h^2}{12} < 10^{-6}$$

resulta que $h < .002\cdots$. O número de intervalos a usar deverá ser $N = \dfrac{\pi/2}{h} \simeq 568.3$, ou seja, pelo menos 569 subintervalos.

Veremos no próximo parágrafo que uma regra de quadratura simples, usando um polinómio interpolador de grau $n = 2$, nos permitirá aproximar o integral com um esforço computacional muito menor. ♦

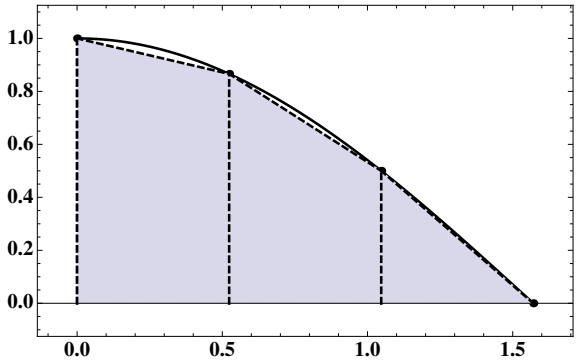


Figura 5.2: O valor aproximado do integral, obtido pela regra dos trapézios composta é igual à soma das áreas dos trapézios assinalados (ver Exemplo 5.3).

5.2 Regra de Simpson

O passo seguinte no estudo da integração numérica consiste em utilizar interpolação quadrática para aproximar a função integranda. Neste caso, para aproximar um dado integral definido

$$I(f) = \int_a^b f(x)dx,$$

precisaremos não de 2, mas de 3 nós no intervalo $[a, b]$. A escolha mais natural destes pontos é $x_0 = a$, $x_1 = (a + b)/2$ e $x_2 = b$, ou seja, o intervalo $[a, b]$ é subdividido em dois subintervalos de igual comprimento $h = (b - a)/2$. Por tal suporte de quadratura passa o polinómio interpolador P_2, de grau ≤ 2. Ao aproximarmos o integral $I(f)$ por

$$Q(f) = \int_a^b P_2(x)\, dx,$$

obtemos uma nova regra de integração numérica conhecida pela designação de *regra de Simpson* simples.

Por construção, a regra de Simpson é da forma

$$S(f) = A_0 f(a) + A_1 f\left(\frac{a+b}{2}\right) + A_2 f(b),$$

onde A_0, A_1 e A_2 são pesos a determinar. Conforme se disse na Secção 5.0.1, os pesos podem ser calculados através da fórmula

$$A_i = \int_a^b L_i(x)\, dx, \quad i = 0, 1, 2\ .$$

Assim, temos

$$A_0 = \int_a^b \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\,dx,$$

$$A_1 = \int_a^b \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\,dx,$$

$$A_2 = \int_a^b \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\,dx\ .$$

Substituídos os valores de x_0, x_1 e x_2, e calculados os integrais anteriores, obtém-se

$$A_0 = \frac{b-a}{6}, \quad A_1 = \frac{4(b-a)}{6}, \quad \text{e} \quad A_2 = \frac{b-a}{6}\ .$$

Por conseguinte, conclui-se que a regra de Simpson simples se escreve

$$\begin{aligned} S(f) = &\quad \frac{b-a}{6}\,f(a) + \frac{4(b-a)}{6}\,f\left(\frac{a+b}{2}\right) + \frac{b-a}{6}f(b) \\ &= \frac{h}{3}\,[f(x_0) + 4\,f(x_1) + f(x_2)], \end{aligned} \tag{5.23}$$

onde $h = (b-a)/2$ representa o comprimento de cada um dos 2 subintervalos anteriormente referidos.

Veremos adiante, no parágrafo 5.3, pág. 258, como calcular mais facilmente os pesos A_i, resolvendo um sistema linear e efectuando uma mudança do intervalo de integração de modo a simplificar os cálculos.

5.2.1 Estimativa de erro na regra de Simpson simples

Uma vez obtida a fórmula (5.23), interessa deduzir uma estimativa do erro a ela associado. Uma questão prévia é a de saber para que polinómios a fórmula de Simpson é exacta.

Por construção, a regra é exacta para qualquer polinómio de grau menor ou igual a 2. Com efeito, se f for um polinómio de grau menor ou igual a 2, então f coincide com o seu polinómio interpolador quadrático, uma vez que, como sabemos, o polinómio interpolador é único. Assim, quaisquer que sejam os 3 nós de interpolação distintos considerados, $f \in \mathcal{P}_2 \Rightarrow I(f) = S(f)$. Por outras palavras, a regra de Simpson é exacta para *qualquer* polinómio de grau menor ou igual a 2.

Além disso, a regra de Simpson oferece-nos um bónus, porquanto ela também é exacta para *polinómios de grau 3*, o que não é tão evidente.

Para nos convencermos disso, comecemos por considerar o caso de $f(x) = x^3$. Sem perda de generalidade, e para facilitar os cálculos, vamos restringir-nos ao intervalo $[-1, 1]$.

Temos que $I(x^3) = \int_{-1}^{1} x^3 dx = 0$. Por outro lado, por aplicação directa da fórmula de Simpson, verifica-se que

$$S(x^3) = \frac{1}{3}(-1)^3 + \frac{4}{3}0 + \frac{1}{3}1^3 = 0 \ .$$

Ou seja, a fórmula de Simpson dá-nos o valor exacto do integral de x^3. Já vimos que a fórmula fornece o valor exacto de qualquer polinómio de grau menor ou igual a 2. Como qualquer polinómio de grau 3 é uma combinação linear de x^3 com um polinómio de grau menor ou igual a 2, somos levados a concluir que *a regra de Simpson é exacta para qualquer polinómio de grau menor ou igual a 3.* Teremos oportunidade de mais adiante chegar à mesma conclusão por outra via.

Esta propriedade traduz uma vantagem da regra de Simpson sobre as outras regras com 3 nós de quadratura. Na realidade, se o terceiro ponto não fosse o ponto médio do intervalo, a regra seria exacta apenas para polinómios de grau menor ou igual a 2. Por isso, ao deduzirmos uma estimativa de erro para a regra de Simpson, devemos preocupar-nos em que essa estimativa de erro reflicta esta propriedade. Isso acontecerá se a estimativa de erro se exprimir através da quarta derivada de f. Nesse caso, se f for um polinómio de grau 3 ou menor, obteremos uma estimativa de erro nula.

Lembremo-nos, a propósito, que no caso da regra dos trapézios a estimativa de erro depende da segunda derivada, o que é coerente com a facto de esta regra ser exacta para polinómios de grau menor ou igual a 1. Como veremos adiante, é possível obter tal estimativa usando as considerações anteriores como guia para o modelo de erro a adoptar em qualquer regra interpolatória.

Tal como fizemos para a regra dos trapézios, comecemos por escrever

$$E_s(f) = I(f) - S(f) = \int_a^b f(x)\,dx - \int_a^b P_2(x)\,dx = \int_a^b (f(x) - P_2(x))dx \ .$$

Considere-se um ponto arbitrário x_3 no intervalo $[a, b]$. É possível construir um polinómio P_3 que interpole f em $x_0 = a$, $x_2 = b$, $x_1 = (a + b)/2$ e também x_3, com x_3 distinto de x_0, x_1 e x_2. Tal polinómio, segundo a fórmula interpoladora de Newton, é dado por

$$P_3(x) = P_2(x) + (x - a)(x - b)(x - (a + b)/2)f[a, b, (a + b)/2, x_3] \ .$$

Verifiquemos que para $x_1 = (a + b)/2$, se tem

$$\begin{aligned} \int_a^b P_3(x)dx &= \int_a^b P_2(x) + f[a, b, x_1, x_3] \int_a^b (x - a)(x - b)(x - x_1)\,dx \\ &= \int_a^b P_2(x)dx, \end{aligned} \tag{5.24}$$

já que é válida a igualdade

$$\int_a^b (x-a)(x-b)(x-(a+b)/2) = 0,$$

devido à simetria do gráfico da função integranda em relação ao ponto médio do intervalo.

De (5.24) resulta,

$$E_s(f) = \int_a^b (f(x) - P_2(x))dx = \int_a^b (f(x) - P_3(x))dx \ . \tag{5.25}$$

Sendo assim, podemos obter uma estimativa de erro para a regra de Simpson a partir da fórmula (5.25), o que nos permite exprimir o erro através da quarta derivada de f, conforme se pretendia.

Apliquemos a fórmula do erro de interpolação. Assumindo que $f \in C^4([a,b])$, temos

$$f(x) - P_3(x) = (x-a)(x-b)(x-x_1)(x-x_3)\,\frac{f^{(4)}(\xi(x))}{4!},$$

para um certo $\xi = \xi(x) \in [a,b]$, donde

$$\begin{aligned} E_s(f) &= \textstyle\int_a^b (f(x) - P_3(x))dx \\ &= \textstyle\int_a^b (x-a)(x-b)(x-x_1)(x-x_3) f^{(4)}(\xi(x))dx \ . \end{aligned} \tag{5.26}$$

Para obter uma estimativa do integral (5.26) iremos recorrer, mais uma vez, ao teorema do valor médio para integrais. No entanto, para isso precisamos de garantir que o polinómio

$$w_4(x) = (x-a)(x-b)(x-x_1)(x-x_3)$$

não muda de sinal no interior do intervalo $[a,b]$. Deveremos portanto especificar um valor adequado de x_3 (o qual até aqui era apenas um ponto arbitrário de $[a,b]$). Na realidade, a *única* maneira de garantir que $w_4(x)$ não muda de sinal no intervalo $[a,b]$ é escolher $x_3 = (a+b)/2 = x_1$. Deste modo, obtém-se

$$w_4(x) = (x-a)(x-b)(x-x_1)^2.$$

Assim, substituindo em (5.26), resulta

$$\begin{aligned} E_s(f) &= \textstyle\int_a^b (f(x) - P_3(x))dx \\ &= \textstyle\int_a^b (x-a)(x-b)(x-x_1)^2 \dfrac{f^{(4)}(\xi(x))}{4!}\, dx \ . \end{aligned} \tag{5.27}$$

Podemos agora aplicar ao integral (5.27) o teorema do valor médio para integrais, considerando

$$u(x) = \frac{f^{(4)}(\xi(x))}{4!}, \quad \text{e} \quad v(x) = w_4(x).$$

Finalmente, obtém-se

$$E_s(f) = \frac{f^{(4)}(\eta)}{4!} \int_a^b (x-a)(x-b)(x-(a+b)/2)^2 dx, \qquad \eta \in (a,b)\ . \tag{5.28}$$

Calculando o integral em (5.28), e fazendo as simplificações necessárias, obtém-se

$$\begin{aligned} E_s(f) = \ & -\left(\frac{b-a}{2}\right)^5 \frac{f^{(4)}(\eta)}{90} = -\frac{h^5}{90} f^{(4)}(\eta) \\ & = -\frac{(b-a)}{180} h^4 f^{(4)}(\eta), \end{aligned} \tag{5.29}$$

uma vez que $h = (b-a)/2$. Por conseguinte, dado que $\eta \in [a,b]$, tem-se

$$\begin{aligned} |E_s(f)| \ & \leq \left(\frac{b-a}{2}\right)^5 \frac{\max_{x\in[a,b]} |f^{(4)}(x)|}{90} \\ & \leq \frac{(b-a)}{180} \max_{x\in[a,b]} |f^{(4)}(x)|\ h^4. \end{aligned} \tag{5.30}$$

A última desigualdade em (5.30) diz-nos que

$$|E_s(f)| = \mathcal{O}(h^4),$$

o que traduz uma grande vantagem relativamente à regra dos trapézios, cujo erro de quadratura é, como sabemos, $\mathcal{O}(h^2)$.

Exemplo 5.4. *Fazendo uma mudança apropriada de variável, pretende-se confirmar a fórmula* (5.29)*, partindo de* (5.28)*.*

Para $h = (b-a)/2$, a função bijectiva

$$\gamma(t) = a + h\,(t+1), \qquad -1 \leq t \leq 1$$

transforma um qualquer ponto $t \in [-1,1]$, num ponto $x \in [a,b]$, e reciprocamente.

Para $x_0 = a$, $x_1 = a+h$ e $x_2 = b = a+2\,h$, designando por K o integral em (5.28), resulta

$$\begin{aligned} K \ & = \int_a^b (x-a)\,(x-x_1)^2\,(x-b)\,dx = \int_{-1}^1 h(t+1) \times (ht)^2 \times h(t-1)\,h\,dt \\ & = h^5 \int_{-1}^1 (t^2-1)t^2\,dt = 2\,h^5 \int_0^1 (t^4-t^2)\,dt \\ & = -\frac{4}{15}\,h^5. \end{aligned}$$

Assim,

$$\begin{aligned} E_s(f) &= -\frac{4}{15\times 4!}\, h^5\, f^{(4)}(\eta) = -\frac{h^5}{90}\, f^{(4)}(\eta) \\ &= -\frac{(b-a)}{180}\, f^{(4)}(\eta)\, h^4, \qquad \eta \in (a,b), \end{aligned}$$

visto que $h = (b-a)/2$. ♦

5.2.2 Regra de Simpson composta

Tal como se fez para a regra dos trapézios composta (ver pág. 247), para aproximarmos o integral $I(f) = \int_a^b f(x)\,dx$, subdivida-se o intervalo $[a,b]$ em N partes. Dado que a regra de Simpson simples utiliza 3 nós, o número $N \geq 2$ deverá ser par.

Fixado $h = (b-a)/N$, em cada um dos subintervalos

$$[x_i, x_{i+2}] = [a + i\,h, a + (i+2)\,h], \quad \text{para} \quad i = 0 : (N-1),$$

é aplicada a regra de Simpson simples (5.23), pág. 251.

A regra de Simpson composta tem como resultado a soma $S_N(f)$ a seguir, a qual se obtém por aplicação da regra simples em cada um dos subintervalos consecutivos $[x_i, x_{i+2}]$. Denotando $f(x_i)$ por f_i, tem-se

$$\begin{aligned} S_N(f) &= \frac{h}{3}\left[(f_0 + 4\,f_1 + f_2) + (f_2 + 4\,f_3 + f_4) + \ldots (f_{N-2} + 4\,f_{N-1} + f_N)\right] \\ &= \frac{h}{3}\left[f_0 + f_N + 4\,(f_1 + f_3 + \ldots + f_{n-1}) + 2\,(f_2 + f_4 + \ldots + f_{N-2})\right]. \end{aligned}$$

Em resumo, sendo $N \geq 2$ par, e $h = (b-a)/N$ o passo da quadratura, a regra de Simpson composta resulta da soma

$$S_N(f) = \frac{h}{3}\left[f(x_0) + f(x_N) + 4\sum_{k=1}^{N/2} f(x_{2\,k-1}) + 2\sum_{k=1}^{N/2-1} f(x_{2\,k})\right]. \tag{5.31}$$

5.2.3 Erro da regra de Simpson composta

Supondo que a função integranda satisfaz a condição $f \in C^4([a,b])$, o erro da regra de Simpson composta obtém-se somando os erros da regra simples cometidos em cada um dos $N/2$ subintervalos $[x_i, x_{i+2}]$, para $i = 0 : (N-2)$.

x_i	$f(x_i)$
0	1.00000000000
$\pi/8 \simeq 0.392699081699$	0.923879532511
$\pi/4 \simeq 0.785398163397$	0.707106781187
$3\,\pi/8 \simeq 1.17809724510$	0.382683432365
$\pi/2 \simeq 1.57079632679$	0

Tabela 5.1: Suporte de quadratura para regra de Simpson composta para 5 nós (ver Exemplo 5.5).

Aplicando a expressão (5.29), resulta

$$\begin{aligned} E_N^S(f) = I(f) - S_N(f) \;\; &= -\frac{h^5}{90}\,\left[f^{(4)}(\eta_1) + f^{(4)}(\eta_2) + \ldots + f^{(4)}(\eta_{N/2})\right] \\ &= -\frac{h^5}{90} \times \frac{N}{2}\,\frac{\left[f^{(4)}(\eta_1) + f^{(4)}(\eta_2) + \ldots + f^{(4)}(\eta_{N/2})\right]}{\frac{N}{2}}, \end{aligned} \tag{5.32}$$

onde $\eta_i \in [x_i, x_{i+2}]$, para $i = 0 : (N-2)$. Visto que por hipótese, a derivada $f^{(4)}$ é contínua em $[a,b]$, atendendo ao teorema do valor intermédio para funções contínuas, existe pelo menos um ponto $\eta \in (a,b)$, para o qual a média aritmética que entra na formação da expressão (5.32) iguala $f^{(4)}(\eta)$, isto é,

$$\begin{aligned} E_N^S(f) = I(f) - S_N(f) \;\; &= -\frac{h^5}{180} \times N\, f^{(4)}(\eta) \\ &= -\frac{b-a}{180}\, h^4\, f^{(4)}(\eta), \qquad \eta \in (a,b), \end{aligned} \tag{5.33}$$

porquanto, $N\,h = b - a$. (Note-se que a fórmula final em (5.33) é formalmente idêntica à expressão (5.29) que deduzimos para a regra de Simpson simples.

Exemplo 5.5. *Pretende-se aproximar o integral*

$$I(f) = \int_0^{\pi/2} \cos(x) dx,$$

(ver Exemplo 5.3, p. 248), mediante aplicação da regra de Simpson composta com:

(a) $N + 1 = 5$, $N + 1 = 9$ *e* $N + 1 = 17$ *nós.*

(b) Calcular E_8/E_4 *e* E_{16}/E_8. *Concluir se os valores numéricos obtidos estão ou não de acordo com a expressão de erro* (5.33).

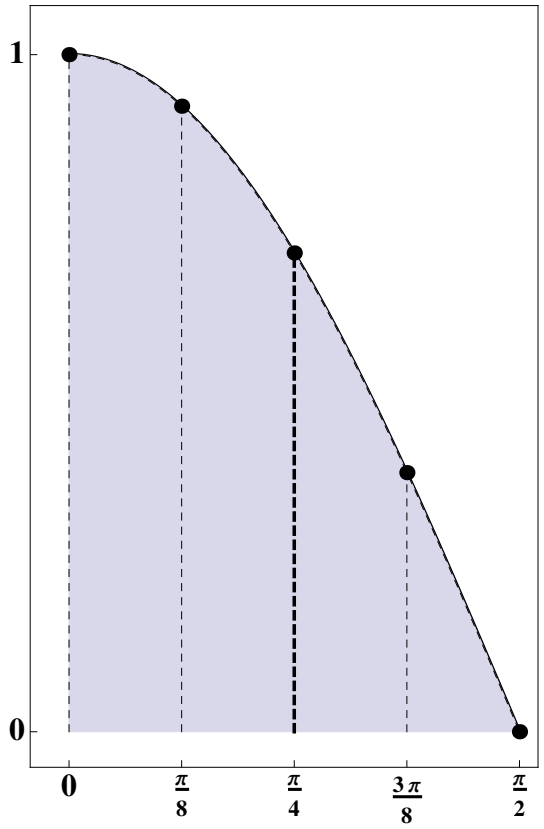


Figura 5.3: Regra de Simpson composta para $N = 4$ subintervalos (Exemplo 5.5).

(a) Para $N = 4$, o passo é $h = \pi/8$. Na Tabela 5.1 encontram-se os valores calculados para o suporte de quadratura (ver Figura 5.3).

Aplicando a fórmula (5.31), obtém-se

$$\begin{aligned} S_4(f) &= \frac{\pi}{24}\left[f(0) + f(\pi/2) + 4\,(\,f(\pi/8) + f(3\,\pi/8)\,) + 2\,f(\pi/4)\,\right] \\ &\simeq 1.000134584974194. \end{aligned}$$

Como $I(f) = 1$, o erro de truncatura é $E_4^S(f) \simeq -0.000134584974194$.

(b) Na Tabela 5.2 encontram-se os valores das aproximações pretendidas e dos respectivos erros.

A fórmula de erro para a regra de Simpson composta, (5.33), diz-nos que

$$|E_N^S(f)| = \mathcal{O}(h^4)\ .$$

Assim, quando passamos de um passo h (ou seja, de um número de subintervalos N), ao passo $h/4$ (ou seja, $2\,N$ subintervalos), o erro da regra reduz-se aproximadamente de $1/16 \simeq 0.0625$. Os valores inscritos na última coluna da Tabela 5.2

N	$S_N(f)$	$I(f) - S_N(f)$	$E_{(2N)/N}$
4	1.000134584974194	-0.000134585	
			0.0616378
8	1.000008295523968	$-8.29552 * 10^-6$	
			0.0622848
16	1.000000516684707	$-5.16685 * 10^-7$	

Tabela 5.2: Comparação de resultados para $N = 4, 8$ e 16.

confirmam esse comportamento do erro de quadratura quando aplicada à função $f(x) = \cos(x)$, no intervalo $[0, \pi/2]$.

♦

5.3 Método dos coeficientes indeterminados

Fixado o inteiro $n \geq 0$, sabemos que uma regra de quadratura interpolatória com $n + 1$ nós, por construção, é exacta para qualquer polinómio de grau $\leq n$.

O Teorema 5.1 a seguir mostra-nos que os pesos de uma regra de quadratura podem ser obtidos resolvendo um determinado sistema de equações lineares. Este processo de cálculo dos pesos recebe a designação de *método dos coeficientes indeterminados*.

Teorema 5.1. Dado $n \geq 0$, a regra de quadratura com $n + 1$ nós distintos

$$Q(f) = A_0\, f(x_0) + A_1\, f(x_1) + \ldots + A_n\, f(x_n),$$

é exacta para qualquer polinómio $p \in \mathcal{P}_n$, se e só se é exacta para todos os elementos de uma base de $\mathcal{P}_n$. Em particular, usando a base canónica de $\mathcal{P}_n$[a] os pesos A_i satisfazem o sistema de equações

$$\begin{aligned} A_0 + A_1 + \cdots + A_n &= \int_a^b dx \\ x_0\, A_0 + x_1\, A_1 + \cdots + x_n\, A_n &= \int_a^b x\, dx \\ &\vdots \\ x_0^n\, A_0 + x_1^n\, A_1 + \cdots + x_n^n\, A_n &= \int_a^b x^n\, dx\ . \end{aligned} \tag{5.34}$$

Além disso, a regra $Q(f)$ é única.

[a] Relembre-se que $\mathcal{P}_n$ designa o espaço linear dos polinómios de grau $\leq n$. A base canónica é constituída pelos monómios $\{1, x, x^2, \ldots, x^n\}$.

Demonstração. Se a regra é exacta para qualquer polinómio $p \in \mathcal{P}_n$, ela é obviamente exacta para os elementos de uma base dos polinómios de grau $\leq n$.

Suponhamos que $\phi_0 = 1, \phi_1, \ldots, \phi_n$ são os elementos de uma base de $\mathcal{P}_n$, e que a regra é exacta para estes elementos, ou seja,

$$Q(1) = I(1),\ Q(\phi_1) = I(\phi_1), \ldots,\ Q(\phi_n) = I(\phi_n)\ .$$

Mostremos que a regra é exacta para qualquer polinómio $p(x) = c_0\,\phi_0(x) + c_1\,\phi_1(x) + \ldots + c_n\,\phi_n(x)$ de $\mathcal{P}_n$.

A regra aplicada ao polinómio p tem a forma

$$\begin{aligned} Q(p) \ &= A_0\,(c_0 + c_1\,\phi_1(x_0) + c_2\,\phi_2(x_0) + \ldots + c_n\,\phi_n(x_0)) + \\ &+ A_1\,(c_0 + c_1\,\phi_1(x_1) + c_2\,\phi_2(x_1) + \ldots + c_n\,\phi_n(x_1)) + \\ &\qquad\vdots \\ &+ A_n\,(c_0 + c_1\,\phi_1(x_n) + c_2\,\phi_2(x_n) + \ldots + c_n\,\phi_n(x_n))\ . \end{aligned}$$

Equivalentemente,

$$\begin{aligned} Q(p) \ &= (A_0 + A_1 + \ldots + A_n)\,c_0 + \\ &+ (A_0\phi_1(x_0) + A_1\,\phi_1(x_1) + \ldots + A_n\,\phi_1(x_n))\,c_1 + \\ &\qquad\vdots \\ &+ (A_0\phi_n(x_0) + A_1\,\phi_n(x_1) + \ldots + A_n\,\phi_n(x_n))\,c_n\ . \end{aligned} \tag{5.35}$$

Ou seja,

$$Q(p) = Q(1)\,c_0 + Q(\phi_1)\,c_1 + \ldots + Q(\phi_n)\,c_n\ .$$

Ora, por hipótese sabemos que $Q(1) = I(1),\ \ldots, Q(\phi_n) = I(\phi_n)$, logo $Q(p) = I(p)$.

Caso a base de $\mathcal{P}_n$ escolhida seja a base canónica, as condições (5.35) traduzem-se no sistema linear (5.34). Uma vez que este sistema possui matriz dos coeficientes que é a transposta da matriz de Vandermonde associada aos nós $x_0, \ldots, x_n$, e sendo estes nós distintos, conclui-se que a matriz é invertível e portanto o sistema (5.34) tem solução única, isto é, a regra de quadratura interpolatória é única. □

Além da base canónica, podem ser utilizadas outras bases para construir o sistema de equações referido no Teorema 5.1. No próximo exemplo veremos como se pode usar a base de Newton, referida na pág. 195, para esse fim.

Exemplo 5.6. *(a) Pretende-se determinar os pesos da regra de quadratura interpolatória, usando os nós* $0, 1, 2$ *e* 3*, para aproximar o integral*

$$I(g) = \int_0^3 g(t)\,dt\ .$$

(b) A partir da regra anteriormente obtida, efectuar uma mudança do intervalo de integração a fim de determinar a expressão da regra correspondente para aproximar o integral

$$I(f) = \int_a^b f(x)\,dx\ .$$

(c)Verificar que a regra que se determinou na alínea (b) é exacta para $\int_0^1 x^3$ *mas não é exacta para* $\int_0^1 x^4$.

(a) Seja

$$Q(g) = A_0\, g(0) + A_1\, g(1) + A_2\, g(2) + A_3\, g(3),$$

a regra cujos pesos pretendemos calcular. Aplicando o método dos coeficientes indeterminados, e utilizando a base de Newton de $\mathcal{P}_3$ associada aos nós dados, ou seja,

$$\phi_0(t) = 1,\ \phi_1(t) = t,\ \phi_2(t) = t\,(t-1), \quad \text{e} \quad \phi_3(t) = t\,(t-1)\,(t-2),$$

obtém-se o seguinte sistema linear triangular superior,

$$\begin{cases} A_0 + A_1 + A_2 + A_3 &= \int_0^3 dt = 3 \\ A_1 + 2\,A_2 + 3\,A_3 &= \int_0^3 t\,dt = 9/2 \\ 2\,A_2 + 6\,A_3 &= \int_0^3 t\,(t-1)\,dt = 9/2 \\ 6\,A_3 &= \int_0^3 t\,(t-1)\,(t-2)\,dt = 9/4\ . \end{cases}$$

A solução deste sistema é $(3/8, 9/8, 9/8, 3/8)^T$. Assim,

$$Q(g) = \frac{3}{8}\,g(0) + \frac{9}{8}\,g(1) + \frac{9}{8}\,g(2) + \frac{3}{8}\,g(3)\ . \tag{5.36}$$

(b) Para $t \in [0,3]$, a função bijectiva $x = \gamma(t) = a + \dfrac{b-a}{3}\,t$, toma valores no intervalo $[a,b]$. Tem-se

$$I(f) = \int_a^b f(x)\,dx = h \int_0^3 f(a + h\,t)\,dt, \quad \text{onde} \quad h = (b-a)/3\ .$$

Sejam $x_0 = a$, $x_1 = a + h$, $x_2 = a + 2\,h$ e $x_3 = b$.

Da mudança de intervalo de integração resulta

$$I(f) = \int_a^b f(x)\,dx = h \int_0^3 f(a + h\,t)\,dt = h \int_0^3 g(t)\,dt\ .$$

Assim,

$$I(f) = h\,I(g), \quad \text{e} \quad Q(f) = h\,Q(g)\ .$$

De (5.36) resulta

$$Q(f) = \frac{3\,h}{8}\,[f(a) + 3\,f(a+h) + 3\,f(a+2\,h) + f(b)], \quad \text{com} \quad h = \frac{b-a}{3}\ . \tag{5.37}$$

(c) Para $f(x) = x^3$, com $x \in [0,1]$, tem-se $h = 1/3$. Logo,

$$Q(x^3) = 1/8\,(0 + 1/3^2 + 2^3/3^2 + 1) = 1/4 = \int_0^1 x^3 dx$$

$$Q(x^4) = 1/8\,(0 + 1/3^3 + 2^4/3^3 + 1) = 11/54 \neq I(x^4)\ .$$

Conclui-se, portanto, que a regra $Q(f)$ é exactamente de grau 3 de exactidão, segundo a Definição 5.1 adiante, pág. 264. ♦

5.3.1 O erro da regra de Simpson revisitado

As regras de quadratura com nós equidistantes num intervalo $[a, b]$ são habitualmente designadas por regras de Newton-Cotes[3].

Uma vez que a regra de Simpson[4] simples utiliza três nós equidistantes (a distância entre nós consecutivos vale $h = (b - a)/2$), trata-se de uma regra de quadratura de Newton-Cotes com 3 nós. A regra (5.37), dita *regra dos* 3/8, é também uma regra de Newton-Cotes com 4 nós.

Fixado o número $n \geq 0$, uma regra de quadratura interpolatória com $n + 1$ nós diz-se *fechada* quando os extremos do intervalo são nós de quadratura. Regras com nós equidistantes em que os extremos a e b do intervalo não são adoptados como nós de quadratura dizem-se regras *abertas*.

Tal como fizemos anteriormente para as regras dos trapézios e de Simpson e no Exemplo (5.6), as regras de Newton-Cotes (fechadas ou abertas) podem facilmente ser obtidas aplicando o método dos coeficientes indeterminados.

Especialmente para $n \geq 2$, a álgebra envolvida é muito facilitada considerando mudanças apropriadas do intervalo de integração e escolhendo uma base de polinómios que facilite os cálculos dos pesos dessas regras. Para esse efeito escolhemos a *base de Newton*, que referimos na página 195.

O erro de quadratura para a regra de Simpson é a seguir deduzido de modo a simplificar os cálculos e *sem recorrer ao teorema do valor médio para integrais*.

Consideremos para intervalo de integração o intervalo $[-1, 1]$, e seja

$$\gamma(t) = a + h\,(t + 1), \quad -1 \leq t \leq 1$$

uma bijecção do intervalo $[-1, 1]$ no intervalo $[a, b]$, tal que

$$g(t) = f(a + h\,(t + 1)) = f(x), \qquad -1 \leq t \leq 1\ . \tag{5.38}$$

Tem-se,

$$I(f) = \int_a^b f(x)\,dx = h \int_{-1}^1 g(t)\,dt\ .$$

Adoptando a notação $Q(\cdot)$ para designar uma regra de quadratura actuando sobre uma determinada função num determinado intervalo, resulta

$$Q(f) = h\,Q(g),$$

onde se subentende que se integra a função f no intervalo $[a, b]$, e g no intervalo $[-1, 1]$.

[3]Roger Cotes, 1682 –1716, matemático inglês, contemporâneo de Newton.

[4]Thomas Simpson, 1710 – 1761, matemático inglês. A regra chamada de Simpson foi usada cerca de 100 anos antes por Johannes Kepler.

Tal como foi mostrado anteriormente, a regra de Simpson simples, para o intervalo $[-1,1]$ tem a forma

$$Q(g) = \frac{1}{3}\left[g(-1) + 4\,g(0) + g(1)\right] .$$

Completando a base de Newton de $\mathcal{P}_2$, $\{1, t+1, (t+1)\,t\}$ (associada aos nós $t_0 = -1, t_1 = 0$ e $t_2 = 1$), de modo a obter uma base de $\mathcal{P}_3$, com um novo elemento $\phi_3(t) \in \Pi_3$[5],

$$\phi_3(t) = (t+1)\,t\,(t-1) = (t^2-1)\,t,$$

(o qual resulta do último elemento da referida base multiplicado por $(t-1)$), concluimos imediatamente que ϕ_3 satisfaz

$$Q(\phi_3) = 0 \quad \text{e} \quad I(\phi_3) = \int_{-1}^{1} \phi_3(t)\,dt = 0 .$$

(Notar que ϕ_3 é função ímpar pelo que $I(\phi_3) = 0$). Assim, por construção, a regra em causa não só é exacta em $\mathcal{P}_2$, mas também para qualquer polinómio de grau ≤ 3.

Tendo em consideração o que se observou a respeito dos erros de quadratura da regra dos trapézios (ver pág. 245) e da regra de Simpson (pág. 251), vamos admitir que no caso da regra de Simpson o respectivo erro possui a forma

$$E_Q(g) = I(g) - Q(g) = c\,g^{(4)}(\theta), \qquad \theta \in (-1,1), \tag{5.39}$$

onde c é uma constante não nula a determinar.

Pretende-se que a fórmula (5.39) seja válida para qualquer função g, pelo menos de classe $C^4([-1,1])$. Em particular que a fórmula referida seja exacta para o polinómio

$$\phi_4(t) = (t+1)^2\,t\,(t-1), \quad \phi \in \Pi_4 .$$

Atendendo a que $Q(\phi_4) = 0$, e $\phi_4^{(4)} = 4!$, substituindo na expressão do erro (5.39), resulta $I(\phi_4) = c \times 4!$, isto é, o valor da constante c do erro de quadratura é

$$c = \frac{I(\phi_4)}{4!},$$

donde

$$E_Q(g) = \frac{I(\phi_4)}{4!}\,g^{(4)}(\theta), \qquad -1 < \theta < 1 . \tag{5.40}$$

Visto que, de (5.38) resulta

$$g^{(4)}(t) = h^4\,f^{(4)}(x),$$

[5] Π_3 denota o conjunto dos polinómios de grau exactamente 3 e coeficiente de maior grau unitário.

a expressão do erro de quadratura para a regra de Simpson aplicada à função f, pode rescrever-se na forma

$$E_S(f) = \frac{I(\phi_4)}{4!}\, h^5\, f^{(4)}(\xi), \quad \xi \in (a,b)\ . \tag{5.41}$$

Dado que

$$I(\phi_4) = \int_{-1}^{1} (t+1)^2\, t\, (t-1)\, dt = \int_{-1}^{1} (t^4 - t^2)\, dt = -\frac{4}{15},$$

substituindo em (5.41), resulta para o erro da regra de Simpson simples,

$$E_S(f) = -\frac{1}{90}\, h^5\, f^{(4)}(\xi), \qquad \xi \in (a,b)\ . \tag{5.42}$$

Atendendo a que no caso da regra de Simpson em $[a,b]$ se tem $h = (b-a)/2$, a expressão anterior pode escrever-se na forma

$$E_S(f) = -\frac{b-a}{180}\, h^4\, f^{(4)}(\xi), \qquad \xi \in (a,b)\ . \tag{5.43}$$

Na expressão anterior o expoente de h é igual à ordem da derivada de f, e evidencia a dependência do erro de quadratura do comprimento $(b-a)$ do intervalo de partida, de acordo com o que já conhecemos (ver (5.29), pág. 254).

Se usarmos a regra de Simpson para integrar uma função polinomial de grau 4, as fórmulas (5.42) e (5.43) permitem obter o valor *exacto* do erro de quadratura $E_S(f)$, já que neste caso a derivada $f^{(4)}$ é constante.

O exemplo a seguir ilustra esse resultado, confirmando heuristicamente que a hipótese formulada imediatamente antes da fórmula (5.39) sobre o comportamento do erro da regra de Simpson, faz todo o sentido.

Exemplo 5.7. *Seja*

$$I(x^4) = \int_{-1}^{1} x^4\, dx = \frac{2}{5},$$

e considere-se a regra de Simpson no intervalo $[-1,1]$*. Neste caso temos* $h = 1$*, e da fórmula* (5.42) *resulta*

$$E_S(x^4) = -\frac{1}{90} \times 1 \times 4! = -\frac{4}{15} = I(x^4) - Q(x^4),$$

isto é, o erro de quadratura é igual a $I(x^4) - Q(x^4)$*, como seria de esperar.*

♦

Deixa-se ao leitor a sugestão para generalizar os argumentos que utilizámos neste parágrafo, a fim de determinar a fórmula e o erro de uma regra de quadratura fechada de Newton-Cotes com 4, 5, 6, ou mais nós.

n	d	A_0	A_1	A_2	A_3	A_4
1	2	1				
2	6	1	4			
3	8	1	3			
4	90	7	32	12		
5	288	19	75	50		
6	840	41	216	27	272	
7	17280	751	3577	1323	2989	
8	28350	989	5888	-928	10496	-4590

Tabela 5.3: Regras de Newton-Cotes fechadas para $1 \leq n \leq 8$.

Fixado $n \geq 1$, para $h = (b-a)/n$, e para os nós $x_i = a + i\,h$, com $i = 0 : n$, as regras de Newton-Cotes fechadas são da forma

$$Q(f) = \frac{b-a}{d} \left(\sum_{i=0}^{n} A_i\, f(x_i) \right).$$

Os pesos A_i são simétricos, isto é, $A_i = A_{n-i}$.

Para $1 \leq n \leq 8$, a Tabela 5.3 indica os pesos e os denominadores d em cada caso. Poderá verificar que para $n \geq 8$ os pesos das respectivas regras de Newton--Cotes deixam de ser todos positivos, o que pode suscitar um comportamento numericamente instável dessas regras no caso em que os seus pesos e nós resultem de valores arredondados. É essa a razão pela qual na prática só são utilizadas regras de Newton-Cotes cujos pesos sejam todos positivos.

5.4 Grau de precisão de regra de quadratura

Vimos que as regras de Newton-Cotes fechadas, com $n+1$ nós num intervalo $[a, b]$, por construção, são exactas para qualquer polinómio de grau $\leq n$.

Mostrámos que a regra de Simpson (para a qual $n = 2$), apresenta a particularidade de ser exacta não apenas para polinómios de $\mathcal{P}_2$, mas também para os polinómios de grau ≤ 3 (prova-se que as regras de Newton-Cotes, com n *par* são exactas para polinómios de $\mathcal{P}_{n+1}$). Por isso se diz que a regra de Simpson tem *grau de precisão* (ou *grau de exactidão*) três, de acordo com a definição a seguir enunciada.

Definição 5.1. Uma regra de quadratura diz-se de grau k, ($k \geq 0$) se e só se é exacta para qualquer polinómio de $\mathcal{P}_k$, mas não é exacta para algum polinómio de grau $k+1$.

Sabemos pelo Teorema 5.1, pág. 258, que o método dos coeficientes indeterminados nos permite obter facilmente os pesos de uma determinada regra de quadratura interpolatória, com $n+1$ nós, aplicando-a aos elementos de uma base qualquer de $\mathcal{P}_n$. Assim, por construção, uma regra de quadratura interpolatória possui grau de exactidão *pelo menos* n.

Como se disse previamente, as regras de Newton-Cotes fechadas, com n par, são regras de grau de exactidão $n+1$. De facto, sabe-se que se escolhermos criteriosamente os nós de quadratura, o grau duma regra pode ser maior do que o que seria previsível levando apenas em conta o grau do polinómio interpolador usado.

O exemplo a seguir ilustra esse facto, com uma regra construída a partir dos *nós de Chebyshev* (ver pág. 209). Trata-se de uma regra *aberta*, visto que os extremos do intervalo de integração não são nós de quadratura.

O mesmo Exemplo 5.8 sugere que algumas regras de quadratura com nós não uniformente distribuidos no intervalo de integração podem ser mais precisas do que as regras com nós equidistantes.

Exemplo 5.8. *(a) Pretende-se determinar os pesos de uma regra de quadratura para aproximar* $I(g) = \int_{-1}^{1} g(t)dt$*, da forma*

$$Q(g) = A_0\, g(t_0) + A_1\, g(t_1) + A_2\, g(t_2),$$

onde os nós são respectivamente os zeros do polinómio de Chebyshev T_3*, referido na página 209:*

$$T_3(t) = 4\,t^3 - 3\,t = t\,(4\,t^2 - 3)\ .$$

(b) Qual é o grau de precisão dessa regra?
(c) Usando como função de teste $g(t) = t^4$*, qual das regras produz um erro de quadratura menor, a regra de Simpson ou a regra obtida na alínea (a)?*

(a) Os zeros do polinómio $T_3(t)$ são

$$t_0 = -\frac{\sqrt{3}}{2}, \quad t_1 = 0, \quad t_2 = \frac{\sqrt{3}}{2}\ .$$

Aplicando o método dos coeficientes indeterminados à base de Newton, seja $\mathcal{N}_3$,

$$\mathcal{N}_3 = \{1, t + \sqrt{3}/2, (t + \sqrt{3}/2)\,t\},$$

resulta o sistema de matriz triangular superior

$$\begin{cases} A_0 + A_1 + A_2 & = \int_{-1}^{1} dt = 2 \\ \dfrac{\sqrt{3}}{2}\,A_1 + \sqrt{3}\,A_2 & = \int_{-1}^{1} (t + \dfrac{\sqrt{3}}{2}) dt = \sqrt{3} \\ \dfrac{3}{2}\,A_2 & = \int_{-1}^{1} (t + \dfrac{\sqrt{3}}{2})\,t dt = \dfrac{2}{3}\ . \end{cases} \tag{5.44}$$

A solução do sistema anterior obtém-se por substituições ascendentes,

$$\begin{aligned} A_2 &= 4/9 \\ A_1 &= (\sqrt{3} - \sqrt{3}\,A_2) \times \frac{2}{\sqrt{3}} = \frac{10}{9} \\ A_0 &= 2 - (A_1 + A_2) = \frac{4}{9}\,. \end{aligned}$$

Por conseguinte, a regra de quadratura tem a forma

$$Q(g) = \frac{1}{9}\left[4\,g\left(-\frac{\sqrt{3}}{2}\right) + 10\,g(0) + 4\,g\left(\frac{\sqrt{3}}{2}\right)\right].$$

(b) Por construção a regra é pelo menos de grau 2. Porém, como

$$Q(t^3) = 0 \quad \text{e} \quad I(t^3) = \int_{-1}^{1} = 0,$$

a regra é pelo menos de grau 3. Mas,

$$Q(t^4) = \frac{1}{2} \quad \text{e} \quad I(t^4) = \frac{2}{5}\,.$$

Assim, como $Q(t^4) \neq I(t^4)$, a regra é exactamente de grau 3.

(c) Apesar da regra anteriormente deduzida ser do mesmo grau que a regra de Simpson, a fórmula de quadratura $Q(g)$ acima pode ser mais interessante. Com efeito, por exemplo, usando como teste a função $g(t) = t^4$, tem-se

$$I(t^4) - Q(t^4) = \frac{2}{5} - \frac{1}{2} = -\frac{1}{10}\,.$$

Ora, uma vez que para a regra de Simpson, $S(g) = \dfrac{1}{3}\,[g(-1) + 4\,g(0) + g(1)]$, o erro para o monómio t^4 é exactamente

$$I(t^4) - S(t^4) = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15},$$

donde se conclui que, neste exemplo, o erro da regra $Q(g)$ é inferior ao erro da regra de Simpson. ♦

5.5 Integrais com função peso *

Nas aplicações são frequentes integrais do tipo

$$I(f) = \int_a^b f(x)\,w(x)\,dx,$$

onde $w(x)$ é uma dada função *não negativa* e integrável em $[a, b]$, habitualmente designada por *função peso*.

No Exemplo 5.9 a seguir, é ilustrado o caso do integral

$$I(g) = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}}\, g(t)\, dt \ . \tag{5.45}$$

A respectiva função peso, $w(t) = (1-t^2)^{-1/2}$, é singular nos pontos ± 1. No entanto, é finito o integral

$$\int_{-1}^{1} \frac{1}{\sqrt{1-t^2}}\, dt = \pi, \tag{5.46}$$

(basta efectuar a mudança de variável $x = \sin(t)$ para nos convencermos que de facto o valor do integral anterior é igual a π).

Uma vez que as regras de Newton-Cotes fechadas, que estudámos anteriormente, utilizam como nós de quadratura os extremos do intervalo de integração, estamos perante um caso em que não é possível construir nenhuma fórmula de Newton-Cotes fechada para aproximar o integral (5.45).

Trata-se de uma situação em que somos naturalmente obrigados a procurar algoritmos alternativos para resolver um problema.

Mostramos no Exemplo 5.9 a seguir, que se reutilizarmos os 3 nós de Chebyshev do Exemplo 5.8, pág. 265 – no contexto actual a fórmula de quadratura que iremos obter é de grau 5 de precisão![6] Alcançamos assim uma vantagem muito substancial comparativamente com a regra de Simpson usada nesse exemplo.

Assim, confirma-se de novo que as regras de quadratura com nós não uniformemente distribuídos podem ser mais vantajosas do que as regras de passo h uniforme.

Exemplo 5.9. *(a) Construir uma regra de quadratura para aproximar o integral* (5.45)*, do tipo*

$$Q(g) = A_0\, g(t_0) + A_1\, g(t_1) + A_2\, g(t_2),$$

uma vez fixados os nós de Chebyshev (ver Exemplo 5.8, pág. 265),

$$t_0 = -\frac{\sqrt{3}}{2}, \quad t_1 = 0, \quad t_2 = \frac{\sqrt{3}}{2} \ .$$

(b) Mostrar que a regra anteriormente obtida é de grau 5 *de precisão.*

[6] Esta regra possui o grau máximo de precisão que é possível obter numa regra interpolatória com 3 nós.

(c) Aplicar a regra $Q(g)$ para calcular exactamente a área assinalada na Figura 5.4, pág 269, ou seja,

$$I = \int_{-1}^{1} \frac{t^6}{\sqrt{1-t^2}}\, dt \ .$$

(d) Dada uma função g integrável, pelo menos de classe $C^6([-1,1])$, obter a fórmula de erro

$$E(g) = I(g) - Q(g),$$

onde $I(g)$ designa o integral (5.45).

(a) Usando o método dos coeficientes indeterminados, para a base de Newton correspondente aos 3 nós de Chebyshev, a matriz do sistema linear resultante é a mesma que se obteve na alínea (a) do Exemplo 5.8, ver (5.44), pág. 265. O segundo membro consiste no vector $(I(1), I(t-t_0), I(t\,(t-t_0))$,

$$I(1) = \int_{-1}^{1} w(t)\, dt = \pi \quad \text{(ver (5.46))},$$

$$I(t-t_0) = \int_{-1}^{1} (t-t_0)\, w(t)\, dt = -t_0 \int_{-1}^{1} w(t)\, dt = \frac{\sqrt{3}}{2}\,\pi,$$

$$I((t-t_0)\,t) = \int_{-1}^{1} (t-t_0)\, t\, w(t)\, dt = \int_{-1}^{1} (t^2 - t\,t_0)\, w(t)\, dt = \frac{\pi}{2} \ .$$

Por conseguinte, o sistema triangular superior a resolver é

$$\left\{ \begin{array}{rl} A_0 + A_1 + A_2 & = \pi \\ \frac{\sqrt{3}}{2}\, A_1 + \sqrt{3}\, A_2 & = \frac{\sqrt{3}}{2}\,\pi \\ \frac{3}{2}\, A_2 & = \frac{\pi}{2}, \end{array} \right. \tag{5.47}$$

de solução $(A_0, A_1, A_2) = (\pi/3, \pi/3, \pi/3)$. Logo,

$$Q(g) = \frac{\pi}{3}\left[g\left(-\frac{\sqrt{3}}{2}\right) + g(0) + g\left(\frac{\sqrt{3}}{2}\right)\right]. \tag{5.48}$$

(b) A base de Newton considerada, $\mathcal{N}_3$, associada aos 3 nós de Chebyshev da quadratura, pode ser estendida a $\mathcal{P}_6$, tal como se indica a seguir.

$$\begin{aligned} \tilde{\mathcal{N}_6} \ &= \{\phi_0(t), \phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t), \phi_5(t), \phi_6(t)\} \\ &= \left\{ 1, t + \frac{\sqrt{3}}{2}, (t+\frac{\sqrt{3}}{2})\, t, (t+\frac{\sqrt{3}}{2})\, t\, (t-\frac{\sqrt{3}}{2}), \right. \\ &\qquad \left. (t+\frac{\sqrt{3}}{2})^2\, t\, (t-\frac{\sqrt{3}}{2}), (t+\frac{\sqrt{3}}{2})^2\, t^2\, (t-\frac{\sqrt{3}}{2}), (t+\frac{\sqrt{3}}{2})^2\, t^2\, (t-\frac{\sqrt{3}}{2})^2 \right\}. \end{aligned}$$

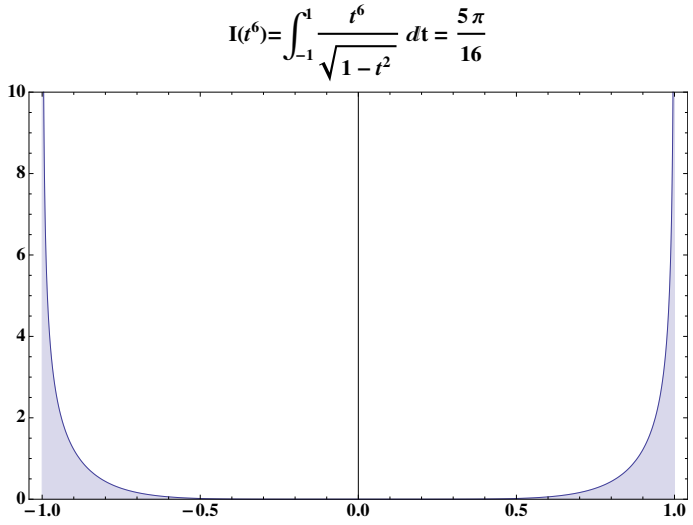


Figura 5.4: A área da região (ilimitada), assinalada a cor, vale $5\,\pi/16$.

Por construção, a regra obtida na alínea anterior é exacta para qualquer polinómio de grau ≤ 2. Além disso, são satisfeitas as igualdades

$$Q(\phi_j) = 0, \quad \text{para} \quad 3 \leq j \leq 6 \ .$$

Dado um polinómio que seja uma função ímpar em $[-1, 1]$, temos simultaneamente $Q(p) = 0$ e $I(p) = \int_{-1}^{1} p(t)\, w(t)\, dt = 0$, uma vez que a função integranda $p\, w$ é ímpar.

Ora, atendendo a

$$I(\phi_4) = \int_{-1}^{1} \left(t^2 - \frac{3}{4}\right)\, t\, w(t)\, dt = 0 \qquad \text{(função integranda ímpar)},$$

e

$$\begin{aligned} I(\phi_5) \ &= \int_{-1}^{1} \left(t - \frac{\sqrt{3}}{2}\right) \left(t + \frac{\sqrt{3}}{2}\right)\, t^2\, w(t)\, dt \\ &= \frac{\sqrt{3}}{2} \int_{-1}^{1} \left(t^2 - \frac{3}{4}\right)\, t^2\, w(t)\, dt = 0, \end{aligned}$$

concluímos que

$$Q(\phi_j) = I(\phi_j), \quad \text{para} \quad 0 \leq j \leq 5 \ .$$

Logo, por construção, a regra é de grau de exactidão pelo menos 5. Pode verificar-se que

$$I(\phi_6) = \int_{-1}^{1} \phi_6(t)\, w(t)\, dt = \frac{\pi}{32} \neq 0,$$

e $Q(\phi_6) = 0$. Assim, a regra é exactamente de grau 5.

(c) Uma vez que a regra possui grau 5 de precisão, admitamos que existe pelo menos um valor θ no intervalo de quadratura, tal que

$$E(g) = I(g) - Q(g) = c\, g^{(6)}(\theta), \quad \theta \in (-1, 1),$$

onde c é uma constante a determinar.

Seja $g(t) = \phi_6(t)$. Sabemos que $Q(\phi_6) = 0$ e $I(\phi_6) = \pi/32$. Logo,

$$E(\phi_6) = \frac{\pi}{32} = c \times 6! \iff c = \frac{\pi}{32 \times 6!} \ .$$

Por conseguinte, a expressão de erro pretendida é

$$E(g) = \frac{\pi}{23\,040}\, g^{(6)}(\theta), \quad \theta \in (-1, 1) \ . \tag{5.49}$$

Note que a técnica aqui seguida para a obtenção da expressão do erro de quadratura, é análoga à que utilizamos no parágrafo 5.3.1 para a dedução do erro da regra de Simpson. ♦

(d) Vamos testar a fórmula de erro (5.49), ensaiando-a com a função polinomial de grau 6, $g(t) = t^6$ (convida-se o leitor a confirmar a validade da expressão de erro adoptada considerando um polinómio qualquer do sexto grau).

Caso $g \in \mathcal{P}_6$, a expressão (5.49) permite-nos calcular *exactamente* o erro de quadratura. Por isso, uma vez calculado $Q(t^6)$, estamos habilitados a calcular exactamente o valor de $I(t^6)$.

Como

$$Q(t^6) = \frac{\pi}{3}\left(2 \times \left(\frac{\sqrt{3}}{2}\right)^6\right) = \frac{9\,\pi}{32},$$

aplicando a igualdade (5.49), resulta

$$I(t^6) - Q(t^6) = \frac{\pi}{32 \times 6!} \times 6! = \frac{\pi}{32} \ .$$

Assim,

$$I(t^6) = \int_{-1}^{1} t^6\, w(t)\, dt = Q(t^6) + \frac{\pi}{32} = \frac{5\,\pi}{16} \ .$$

Pode verificar-se (tal como é indicado na Figura 5.4) que de facto o valor de $I(t^6)$ é o obtido na expressão anterior, confirmando a consistência do modelo de erro de quadratura utilizado.

5.6 Regras compostas *

Uma regra de quadratura-padrão, habitualmente designada "simples" (tal como a regra dos trapézios ou de Simpson simples), pode ser aplicada sucessivamente numa partição de um intervalo $[a, b]$. Somando os valores obtidos temos uma *regra composta*, de que são exemplos a regra dos trapézios composta, discutida no parágrafo 5.1.2, pág. 246 , ou a regra de Simpson composta, de que nos ocupámos no parágrafo 5.2.2, pág. 255.

Num contexto mais geral, para aproximar o integral

$$I(g) = \int_{\alpha}^{\beta} g(t)\, w(t)\, dt,$$

(onde w é uma função peso dada), vamos admitir termos já construído uma determinada regra de quadratura-padrão, seja

$$Q(g) = A_0\, g(t_0) + A_1\, g(t_1) + A_2\, g(t_2)\ .$$

São muito comuns "intervalos-padrão"como $[\alpha, \beta] = [-1, 1]$ ou $[\alpha, \beta] = [0, 1]$, ou outros para os quais a função peso w possui certas propriedades interessantes para as aplicações.

Em geral pretende-se calcular aproximações $I(f) = \int_a^b f(x)\, w(x)\, dx$, pelo que deveremos relacionar o cálculo de uma aproximação de quadratura $Q(f)$ no intervalo $[a, b]$, com a aproximação $Q(g)$ no intervalo $[\alpha, \beta]$. A ideia é aplicar a fórmula de quadratura padrão sucessivamente num certo número de subintervalos de $[a, b]$.

Para ilustrarmos o procedimento, apenas lidaremos com regras de 3 nós, $t_i \in [\alpha, \beta]$, mas as considerações a seguir são facilmente generalizáveis para uma regra-padrão com qualquer outro número de nós.

Ao contrário das regras dos trapézios e de Simpson compostas, anteriormente estudadas, nas fórmulas que iremos deduzir nesta secção, o espaçamento entre nós de uma regra composta poderá ser qualquer.

Nesse sentido, é útil designar por *célula computacional*, qualquer intervalo

$$[x_i, x_{i+1}] \subseteq [a, b],$$

onde $[a, b]$ é um intervalo onde será construída a regra de quadratura composta associada à regra-padrão de partida.

Designemos por $h_i = x_{i+1} - x_i$, o comprimento de uma célula computacional $[x_i, x_{i+1}]$. Vejamos como reescrever a fórmula $Q(g)$ quando aplicada numa dada célula computacional, ou seja $Q(g_i)$.

Comecemos por definir a bijecção γ_i a seguir, na qual um ponto genérico t do intervalo $[\alpha, \beta]$ é transformado no ponto x da célula computacional,

$$x = \gamma_i(t) = x_i + \frac{h_i}{\beta - \alpha}\,(t - \alpha), \qquad t \in [\alpha, \beta]\ .$$

Por conseguinte, aos nós t_0, t_1 e t_2 da regra-padrão correspondem os seguintes nós da célula computacional:

$$\begin{aligned} z_{0,i} &= x_i + \frac{h_i}{\beta - \alpha}\,(t_0 - \alpha) \\ z_{1,i} &= x_i + \frac{h_i}{\beta - \alpha}\,(t_1 - \alpha) \\ z_{2,i} &= x_i + \frac{h_i}{\beta - \alpha}\,(t_2 - \alpha)\ . \end{aligned} \tag{5.50}$$

Fazendo

$$g_i(t) = f(\gamma_i(t)) = f(\, x_i + \frac{h_i}{\beta - \alpha}\,(t - \alpha)\,),$$

e atendendo a que

$$I(f_i) = \int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{h_i}{\beta - \alpha}\int_{\alpha}^{\beta} f(\gamma_i(t))\,dt, \tag{5.51}$$

temos,

$$I(f_i) = \int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{h_i}{\beta - \alpha}\,I(g_i)\ . \tag{5.52}$$

Logo,

$$Q(f_i) = \frac{h_i}{\beta - \alpha}\,Q(g_i)\ . \tag{5.53}$$

Estamos agora habilitados a construir a regra de quadratura composta no intervalo $[a, b]$, somando as regras construídas em cada célula computacional.

Com efeito, se no intervalo $[a, b]$ considerarmos uma partição com $N\,(N \geq 1)$ células computacionais c_i, com

$$c_1 = [x_0, x_1],\ c_2 = [x_1, x_2], \ldots, c_n = [x_{N-1}, x_N],$$

e sendo $Q(g_i)$ a correspondente regra para cada célula, tem-se

$$\begin{aligned} Q(g_i) \quad &= A_0\,g_i(t_0) + A_1\,g_i(t_1) + A_2\,g_i(t_2) \\ &= \frac{h_i}{\beta - \alpha}\,(A_0\,f(z_{0,i}) + A_1\,f(z_{1,i}) + A_2\,f(z_{2,i})) =, \frac{h_i}{\beta - \alpha}\,Q(f_i), \qquad i = 1 : N\ . \end{aligned} \tag{5.54}$$

A regra composta é

$$Q^N(f) = \sum_{i=1}^{N} Q(f_i) = \frac{1}{\beta - \alpha}\,\sum_{i=1}^{N} h_i\ (A_0\,f(z_{0,i}) + A_1\,f(z_{1,i}) + A_2\,f(z_{2,i}))\ . \tag{5.55}$$

Exemplo 5.10. *Pretende-se obter uma aproximação de*

$$I(f) = \int_0^{\pi/2} \cos(x)\, dx$$

(Ver Exemplo 5.3, p. 248).

(a) Considere para regra padrão a regra com 3 nós de Legendre (ver adiante (5.64)*, pág. 282), definida em* $[\alpha, \beta] = [-1, 1]$*, dada por*

$$Q(g) = \frac{1}{9}\,(5\, g(t_0) + 8\, g(t_1) + 5\, g(t_2)),$$

onde $t_0 = -\sqrt{3/5}$*,* $t_1 = 0$*,* $t_2 = \sqrt{3/5}$*. Obter a correspondente regra composta, usando células computacionais de passo uniforme* $h = \pi/6$ *(ou seja, considerando* $N = 3$ *subintervalos de igual comprimento* $h = \pi/6$*, em* $[a, b] = [0, \pi/2]$*.*

(b) Comparar o erro da regra composta anterior com o erro calculado na pág. 243 para a regra dos trapézios composta e com o erro para a regra de Simpson (ver pág. 256).

(a) No intervalo $[a, b] = [0, \pi/2]$, subdividido em $N = 3$ partes, de comprimento $h = \pi/6$, considerem-se as células computacionais,

$$c_i = [x_i, x_{i+1}] = [i\, h, (i + 1)\, h], \qquad i = 1 : 3\ .$$

Em cada célula ficam definidos os 3 nós correspondentes aos nós de Legendre da regra-padrão,

$$z_{0,i} = x_i + \frac{h}{2}\,(t_0 + 1)$$

$$z_{1,i} = x_i + \frac{h}{2}\,(t_1 + 1) = x_i + \frac{h}{2} \qquad i = 1 : 3$$

$$z_{2,i} = x_i + \frac{h}{2}\,(t_2 + 1)\ .$$

A respectiva regra composta é

$$Q_N(f) = \frac{h}{18} \sum_{i=1}^{3} 5\, f(z_{0,i}) + 8\, f(z_{1,i}) + 5\, f(z_{2,i})\ .$$

(b) Aplicando a fórmula anterior, obtém-se,

$$Q_N(f) = 1.00000001071725\ .$$

Como $I(f) = 1$, o erro de truncatura é $E_{Q_N}(f) = |I(f) - Q_N(f)| \simeq 1.07 \times 10^{-8}$, muito inferior a 0.0229, que é o erro cometido quando aplicamos as regras compostas dos trapézios e de Simpson (ver pág. 250 e pág. 256). ⧫

5.7 Exercícios resolvidos

Exercício 5.1. *Sendo dada a equação diferencial*

$$y'(x) = \cos(x), \qquad 0 \leq x \leq \pi/2,$$

tal que $y(0) = 0$*, pretende-se estimar o valor da solução* $y(x) = \sin(x)$ *nos pontos*

$$x_i \in \{0, \pi/8, \pi/4, 3\,\pi/8, \pi/2\}\,,$$

aplicando a regra de Simpson.

(a) Obtenha uma tabela $\{(x_i, y_i)\}_{i=0}^{i=4}$*, onde* y_i *designa uma aproximação da solução do problema*[7] *em cada ponto* x_i*, aplicando a regra de Simpson composta (reveja o Exemplo 5.5 pág. 256).*

Em cada caso deverá ajustar o número de subdivisões N *do intervalo de quadratura em causa de modo a usar sempre o mesmo passo, de valor* $h = \pi/16$*.*

(b) Construa o polinómio $p_4(x)$*, interpolador da tabela que obteve na alínea anterior.*

Compare a gráfico do erro de interpolação $e_4(x) = \sin(x) - p_4(x)$*, com o gráfico da Figura. 5.5, e conclua a respeito do número de algarismos significativos que poderia garantir para um valor de* $y(x) \simeq p_4(x)$*, para* $0 \leq x \leq \pi/2$*, caso usasse o polinómio interpolador como aproximação da solução* $y(x) = \sin(x)$ *do problema de valor inicial dado.*

(a) Fixado um ponto $x_i \in [0, \pi/2]$, integrando ambos os membros da equação diferencial dada, tem-se

$$y(x_i) = y(0) + \int_0^{x_i} \cos(x)\, dx = \int_0^{x_i} \cos(x)\, dx\ .$$

Assim, se substituirmos o integral na expressão anterior por uma sua aproximação, calculada mediante uma determinada regra de quadratura, obtemos uma aproximação y_i da solução exacta $y(x_i)$ do problema de valor inicial dado. O erro de truncatura cometido será, portanto, o erro da quadratura utilizada.

Relembre-se que no Exemplo 5.5, pág. 256, foi calculado o valor $y_4 \simeq y(\pi/2)$ pela regra de Simpson composta, usando um passo de quadratura $h = \pi/16 \simeq 0.19635$ (para $N = 8$ subintervalos de $[0, \pi/2]$.

Dado que, como sabemos, para um intervalo $[a, b]$, subdividido em N partes de comprimento $h = (b - a)/N$, o erro de quadratura para a regra de Simpson composta é da ordem $\mathcal{O}(h^4)$, o erro de quadratura do valor y_4 mencionado será da ordem de $(\pi/16)^4 \simeq 0.0015$.

[7] Problemas do tipo proposto dizem-se problemas de valor inicial. Métodos numéricos para a sua resolução serão discutidos no Capítulo 6.

x_i	$N = \lceil x_i/h \rceil$	$y_i = S_N(\cos(x))$	Erro de quadratura
0		0	0
$\pi/8$	2	0.3826866069246750	$-3.17456 * 10^{-}6$
$\pi/4$	4	0.7071126470077986	$-5.86582 * 10^{-}6$
$3\,\pi/8$	6	0.9238871965760920	$-7.66406 * 10^{-6}$
$\pi/2$	8	1.000008295523968	$-8.29552 * 10^{-6}$

Tabela 5.4: Regra de Simpson composta para $N + 1$ nós.

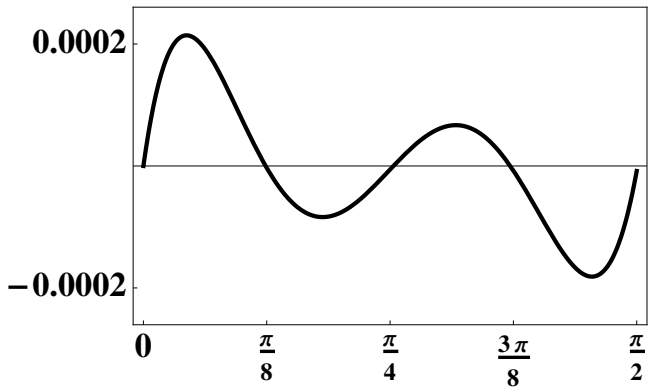


Figura 5.5: Erro de interpolação $e_4(x) = \sin(x) - p_4(x)$.

De modo análogo, para calcularmos a tabela de valores pedida, iremos ajustar o número de subintervalos N, de modo a garantir, em cada caso, um erro de quadratura não superior a 0.0015, quando aplicada a regra de Simpson no intervalo em causa (respectivamente $[0, \pi/8]$, $[0, \pi/4]$, $[0, 3\,\pi/8]$ e $[0, \pi/2]$).

Na Tabela 5.4 mostram-se os resultados obtidos. Adoptado o passo comum $h = \pi/16$, o número N de subdivisões de cada intervalo $[0, x_i]$ é[8] $N = \lceil x_i/h \rceil$. O símbolo $S_N(\cos x)$ indica o valor calculado pela regra de Simpson para o correspondente valor de N.

Conforme se pode verificar na última coluna da Tabela 5.4, por exemplo o erro de quadratura para $y_1 \simeq y(\pi/8)$ é da ordem de 10^{-6}, muito inferior ao que grosseiramente se poderia antever apenas através da expressão $\mathcal{O}(h^4)$ (para $h = \pi/16$ é $h^4 \simeq 0.001486$).

De facto, aplicando a fórmula de majoração de erro (5.33), pág. 256, para $h = \pi/16$, e uma vez que $|\cos^{(4)}(x)| \leq 1,\ \forall x \in [0, \pi/8]$, temos

$$|I(\cos(x)) - S_2(\cos(x))| \leq \frac{\pi/8}{180} \times \left(\frac{\pi}{16}\right)^4 \simeq 3.2 \times 10^{-6},$$

[8] A função inteira "ceiling", de símbolo $\lceil x \rceil$, dá o menor inteiro não inferior ao número real x.

resultado que está de acordo com o respectivo erro de quadratura tabelado.

(b) Usando uma das fórmulas de interpolação que estudou no Capítulo 4, podemos calcular o seguinte polinómio $p_4(x)$, interpolador dos valores (x_i, y_i) da Tabela 5.4,

$$\begin{aligned} p_4(x) \quad &= 0.99632524358504\, x + 0.01995159501150\, x^2 - 0.20358714963439\, x^3 + \\ &+ 0.02871446342973\, x^4 \ . \end{aligned}$$

Por inspecção do gráfico da Figura 5.5, onde está traçada a função erro de interpolação para a solução do problema de valor inicial, $e_4(x) = y(x) - p_4(x)$, conclui-se que o erro absoluto máximo de interpolação é aproximadamente de 0.0002, pelo que qualquer estimativa da solução $y(x) = \sin(x)$, no intervalo $[0, \pi/2]$, através de $p_4(x)$, terá pelo menos 3 algarismos significativos. ♦

Algumas regras de Newton-Cotes abertas

As regras de Newton-Cotes fechadas não são aplicáveis quando a função integranda não está definida em um ou em ambos os extremos do intervalo de integração. Por exemplo, não podemos usar a regra dos trapézios para aproximar o integral

$$I(f) = \int_a^b f(x)\, dx = \int_0^{1/2} \frac{\sin(x)}{x}\, dx, \tag{5.56}$$

uma vez que a função integranda não está definida em $x = 0$. No entanto, as regras abertas poderão ser utilizadas para calcular (5.56).

As regras de Newton-Cotes abertas, com apenas um nó, são respectivamente conhecidas pela designação de regra do *rectângulo à esquerda*, *rectângulo à direita* e regra do *ponto médio*. Fazendo $h = b - a$ e designando por $L(f)$, $R(f)$ e $M(f)$ as referidas regras, tem-se

$$\begin{aligned} L(f) &= h\, f(a) \qquad \text{(rectângulo à esquerda)} \\ R(f) &= h\, f(b) \qquad \text{(rectângulo à direita)} \\ M(f) &= h\, f\left(\frac{a+b}{2}\right) \qquad \text{(ponto médio).} \end{aligned} \tag{5.57}$$

Caso f seja uma função positiva, cada uma das expressões anteriores representa a área de um rectângulo, o que justifica a designação dada à regras mencionadas. As regras (5.57) podem ser usadas nomeadamente para aproximar a solução de uma equação diferencial, tal como é referido no Capítulo 6, pág. 290.

Supondo que a função integranda é suficientemente regular, pode usar-se o método dos coeficientes indeterminados (ver Exercício 5.2 adiante) para obter as seguintes

expressões do erro dessas regras (5.57):

$$\begin{aligned} E_L(f) &= \frac{b-a}{2}\, f'(r)\, h, \qquad r \in (a,b) \\ E_R(f) &= -\frac{b-a}{2}\, f'(s)\, h, \qquad s \in (a,b) \\ E_M(f) &= \frac{b-a}{24}\, f^{(2)}\xi)\, h^2, \qquad \xi \in (a,b) \end{aligned} \tag{5.58}$$

As expressões de erro anteriores traduzem o facto das regras $L(f)$ e $R(f)$ serem de grau zero de precisão, enquanto a regra $M(f)$ é de grau um.

Se a função f' não mudar de sinal em $[a,b]$, conclui-se de (5.58) que o erro de quadratura de $L(f)$ tem sinal contrário ao erro de $R(f)$, donde as majorações de erro,

$$\begin{aligned} |I(f) - L(f)| &\leq |L(f) - R(f)| \\ |I(f) - R(f)| &\leq |L(f) - R(f)| \,. \end{aligned} \tag{5.59}$$

Supondo que a função f' é constante no intervalo de integração, resulta de (5.58) que

$$I(f) - L(f) = -(I(f) - R(f)) \Longleftrightarrow I(f) = \frac{L(f) + R(f)}{2} \,.$$

Assim, no caso geral em que a função f' não é constante, o membro direito da última igualdade aproxima $I(f)$. Designemos por $T(f)$ essa aproximação:

$$T(f) = \frac{L(f) + R(f)}{2} = \frac{h}{2}\, (f(a) + f(b)) \,.$$

Ou seja, obtém-se o mesmo resultado da regra dos trapézios, a qual pode ser considerada como a *média aritmética* das regras do rectângulo à esquerda e à direita.

Do mesmo modo que as regras do rectângulo à esquerda e à direita estão relacionadas com a regra dos trapézios, vejamos como relacionar a regra do ponto médio com a regra de Simpson.

Supondo que $f^{(2)}$ não muda de sinal em $[a,b]$, atendendo a que o erro da regra dos trapézios tem por expressão $E_T(f) = -\dfrac{b-a}{12}\, f^{(2)}(\eta)\, h^2$ (ver pág 245), conclui-se de (5.58) que o erro da regra dos trapézios tem sinal oposto ao do erro da regra do ponto médio. Por conseguinte, sob a hipótese referida sobre f'', tem-se

$$|I(f) - M(f)| \leq |T(f) - M(f)| \,. \tag{5.60}$$

Admitindo que $f^{(2)}$ é contante no intervalo de integração, resulta de (5.58) que

$$I(f) - M(f) = -\left(I(f) - T(f)\right)/2 \Longleftrightarrow I(f) = \frac{T(f) + 2\,M(f)}{3}\ .$$

No caso geral, a última igualdade dá-nos uma aproximação de $I(f)$, seja $S(f)$, com

$$S(f) = \frac{T(f) + 2\,M(f)}{3} = \frac{b-a}{6}\ \left(f(a) + f(b) + 4\,f\left(\frac{a+b}{2}\right)\right).$$

O resultado é o mesmo que o da regra de Simpson. Por conseguinte, esta regra pode ser considerada como uma *média pesada* das regras do trapézio e do ponto médio.

Exercício 5.2. *Considere-se o integral* $I(f) = \int_a^b f(x)\,dx$, *onde* $f \in C^2[(a,b)]$.

(a) Deduzir as expressões de erro (5.57) *para as regras do rectângulo à esquerda, do rectângulo à direita e do ponto médio.*

(b) Calcular o integral (5.56), *mediante aplicação da regra do ponto médio composta, com erro não superior a* $\epsilon = 10^{-4}$.

(c) A partir do desenvolvimento de Taylor

$$\frac{\sin(x)}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \mathcal{O}\left(x^8\right),$$

obter uma aproximação do integral (5.56) *com erro inferior a* 10^{-6}.

(a) Para deduzirmos as fórmulas de erro pretendidas, comecemos por considerar o intervalo $[-1,1]$. Iremos aplicar o método dos coeficientes indeterminados neste intervalo, usando a base polinomial $\phi_0(t) = 1$, $\phi_1(t) = t - t_0$ e $\phi_2(t) = (t-t_0)\,(t-t_1)$, onde os nós t_i serão fixados em função da regra de quadratura a tratar.

Sejam $h = b - a$ e $x = \gamma(t) = a + \dfrac{h}{2}\,(t+1)$ a bijecção linear que leva o intervalo $[-1,1]$ no intervalo $[a,b]$. Tem-se,

$$\int_a^b f(x)dx = \frac{h}{2}\int_{-1}^{1} g(t)dt, \quad \text{com} \quad g(t) = f(x) = f(a+\frac{h}{2}\,(t+1)), \quad -1 \le t \le 1\ .$$

Erro da regra do rectângulo à esquerda.

$$\begin{aligned} &t_0 = -1 \Longrightarrow \phi_1(t) = t + 1 \\ &L(g) = 2\,g(-1) \Longrightarrow L(f) = \frac{h}{2}\,L(g) = h\,f(a)\ . \end{aligned}$$

Como $L(\phi_1) = 0$ e $I(\phi_1) = \int_{-1}^{1} t + 1\,dt = 2$, resulta

$$E_L(g) = I(\phi_1)\,g'(\theta), \quad \theta \in (-1,1),$$

e

$$E_L(f) = I(f) - L(f) \;\; = \frac{h}{2}\, E_L(g) = \frac{h}{2} I(\phi_1) \;\times \frac{h}{2}\, f'(r), \quad r \in (a,b)$$

$$= \frac{h^2}{2}\, f'(r) = \frac{b-a}{2} f'(r)\, h \;.$$

Erro da regra do rectângulo à direita.

$$\begin{aligned} & t_0 = 1 \Longrightarrow \phi_1(t) = t - 1 \\ & R(g) = 2\, g(1) \Longrightarrow R(f) = \frac{h}{2}\, R(g) = h\, f(b) \;. \end{aligned}$$

Como $R(\phi_1) = 0$ e $I(\phi_1) = \int_{-1}^{1} t - 1\, dt = -2$, resulta

$$E_R(g) = I(\phi_1)\, g'(\theta), \quad \theta \in (-1,1),$$

e

$$E_R(f) = I(f) - R(f) \;\; = \frac{h}{2}\, E_R(g) = \frac{h}{2} I(\phi_1) \;\times \frac{h}{2}\, f'(s), \quad s \in (a,b)$$

$$= -\frac{h^2}{2}\, f'(s) = -\frac{b-a}{2} f'(s)\, h \;.$$

Erro da regra do ponto médio.

$$\begin{aligned} & t_0 = 0 \Longrightarrow \phi_1(t) = t \\ & t_1 = 1 \Longrightarrow \phi_2(t) = t\,(t-1) \\ & M(g) = 2\, g(0) \Longrightarrow M(\phi_1) = M(\phi_2) = 0 \;. \end{aligned}$$

Como $I(\phi_2) = \int_{-1}^{1} t\,(t-1)\, dt = \frac{2}{3}$, tem-se

$$E_M(g) = \frac{I(\phi_2)}{2!}\, g^{(2)}(\theta), \quad \theta \in (-1,1),$$

e

$$E_M(f) = I(f) - M(f) \;\; = \frac{h}{2}\, E_M(g)$$

$$= \frac{h}{2} \times \frac{1}{3} \times \left(\frac{h}{2}\right)^2 f^{(2)}(\xi), \quad \xi \in (a,b) \tag{5.61}$$

$$= \frac{h^3}{24}\, f^{(2)}(\xi) = \frac{b-a}{24}\, f^{(2)}(\xi)\, h^2.$$

Regra do ponto médio composta.

Subdividindo o intervalo $[a, b]$ em $N \geq 1$ partes de comprimento $h = (b-a)/N$, considerem-se os N nós,

$$x_i = a + (2\,i - 1)\,\frac{h}{2}, \qquad i = 1 : N. \tag{5.62}$$

A regra do ponto médio composta escreve-se

$$M_N(f) = h \sum_{i=1}^{N} f(x_i), \tag{5.63}$$

onde os nós de quadratura são dados por (5.62).
Deixa-se ao leitor a dedução da expressão do erro da regra do ponto médio composta, a qual é idêntica à que se obteve em (5.61), fazendo $h = (b-a)/N$.

(b) A função $f(x) = \sin(x)/x$ e as suas derivadas podem prolongar-se por continuidade ao intervalo $[0, 1/2]$. Tem-se, para $x \in (0, 1/2]$,

$$\begin{aligned}
f'(x) &= \frac{x\,\cos(x) - \sin(x)}{x^2} < 0 \quad \text{e} \quad \lim_{x\to 0^+} f'(x) = 0 \\
f^{(2)}(x) &= \frac{(x^2-2)\,\sin(x) + 2\,x\,\cos(x)}{x^3} < 0 \quad \text{e} \quad \lim_{x\to 0^+} f^{(2)}(x) = -1/3 \\
f^{(3)}(x) &= \frac{3\,(x^2-2)\,\sin(x) - x\,(x^2-6)\,\cos(x)}{x^4} > 0 \quad \text{e} \quad \lim_{x\to 0^+} f^{(3)}(x) = 0\ .
\end{aligned}$$

Assim, a função $f^{(2)}$ é negativa e crescente no intervalo $[0, 1/2]$. Seja

$$M = \max_{0\leq x\leq 1/2} |f''(x)| = |f^{(2)}(0)| = 1/3\ .$$

Vamos determinar o número de subintervalos N do intervalo $[0, 1/2]$, de modo que a regra (5.63) possua um erro não superior ao valor ϵ dado. Como $b-a = 1/2$, de (5.61) obtém-se

$$\frac{1}{48} \times \frac{1}{3} \times \frac{1}{(2\,N)^2} \leq \epsilon \Longleftrightarrow N \geq \sqrt{\frac{1}{576\,\epsilon}} = \frac{25}{6} \simeq 4.2\ .$$

Fixando $N = 5$, isto é, $h = 1/10$, a aproximação pretendida é

$$M_5(f) = \frac{1}{10}\,[f(1/20) + f(3/20) + f(5/20) + f(7/20) + f(9/20)] = 0.493175\ .$$

(b) Uma vez que para $x \in (0, 1/2]$ a série de Taylor da função $\sin(x)/x$ é alternada e de termos absolutamente decrescentes, tem-se que se retivermos os 4 primeiros

termos do desenvolvimento, o respectivo erro é menor do que o erro absoluto do primeiro termo desprezado, isto é,

$$\left| \frac{\sin(x)}{x} - \sum_{i=0}^{3} (-1)^i \frac{x^{2\,i}}{(2\,i+1)!} \right| < \frac{x^8}{9!} \leq (1/2)^8/9! < 10^{-7}.$$

Assim,

$$I(f) \simeq \int_0^{1/2} 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!}\, dx = 0.4931074174,$$

com erro inferior a 10^{-6}. ♦

Exercício 5.3. *Dado o integral*

$$I(f) = \int_{-1}^{1} f(x)\, dx,$$

pretende-se construir uma fórmula para o aproximar, da forma

$$Q(f) = A_0 f(0) + A_1 \left[f(x_1) + f(-x_1) \right], \qquad x_1 \neq 0 \ .$$

(a) É possível escolher o nó x_1 de modo que a regra de quadratura possua exactamente grau 5 de precisão? No caso positivo obtenha essa fórmula.

(b) Os polinómios de Legendre podem ser definidos recursivamente ([12], p. 462, [10], p. 198) pelas expressões

$$\begin{aligned} &P_0(x) = 1 \\ &P_1(x) = x \\ &P_{n+1}(x) = \frac{2\,n+1}{n+1}\, x\, P_n(x) - \frac{n}{n+1}\, P_{n-1}(x), \qquad n = 1, 2, \ldots \ . \end{aligned}$$

Verifique que os três nós da regra que determinou na alínea anterior são zeros do polinómio de Legendre do terceiro grau.

(c) Para aproximar

$$I(f) = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\, dx = \pi,$$

a fórmula que determinou na alínea (a) é mais precisa do que a fórmula de quadratura que se deduziu no Exemplo 5.9, pág. 267, a qual utiliza três nós de Chebyshev? Justifique.

(a) Aplicando o método dos coeficientes indeterminados para a base canónica[9], tem-se

$$\begin{cases} A_0 + 2\, A_1 = \int_1^1 dx = 2 \\ \qquad 2\, x_1^2\, A_1 = \int_{-1}^1 x^2\, dx = \dfrac{2}{3} \ . \end{cases}$$

[9]Se refizer os cálculos partindo da base de Newton associada aos nós dados, deverá obter o mesmo resultado, uma vez que a regra de quadratura interpolatória é única.

Logo,

$$A_1 = \frac{1}{3\,x_1^2}, \quad A_0 = 2 - 2\,A_1 = \frac{6\,x_1^2 - 2}{3\,x_1^2}\ .$$

Assim, por construção, a fórmula a seguir é de grau 2 de exactidão (pelo menos):

$$Q(f) = \frac{1}{3\,x_1^2}\,f(-x_1) + \frac{6\,x_1^2 - 2}{3\,x_1^2}\,f(0) + \frac{1}{3\,x_1^2}\,f(x_1)\ .$$

Uma vez que para qualquer polinómio do tipo $p(x) = x^k$, com k ímpar, se tem $Q(x^k) = I(x^k) = 0$, então a regra é pelo menos de grau 3 de precisão.

Vamos de seguida determinar um valor do nó x_1, de modo que a regra seja pelo menos de grau 4.

$$Q(x^4) = I(x^4) \iff \frac{2\,x_1^4}{3\,x_1^2} = \frac{2}{5} \iff x_1 = \pm\sqrt{\frac{3}{5}}\ .$$

Por conseguinte, escolhido $x_1 = \sqrt{\dfrac{3}{5}}$, visto que $Q(x^5) = I(x^5)$, a regra será pelo menos de grau 5. Como $Q(x^6) = \dfrac{6}{25} \neq I(x^6)$, então a seguinte regra é exactamente de grau 5,

$$\begin{aligned} Q(f) \ &= \frac{5}{9}\,f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}\,f(0) + \frac{5}{9}\,f\left(\sqrt{\frac{3}{5}}\right) \\ &= \frac{1}{9}\left[5\,f\left(-\sqrt{\frac{3}{5}}\right) + 8\,f(0) + 5\,f\left(\sqrt{\frac{3}{5}}\right)\right]. \end{aligned} \tag{5.64}$$

(b) O polinómio de Legendre, de grau 3, é

$$P_3(x) = \frac{1}{2}\,x\,(5\,x^2 - 3),$$

cujos zeros coincidem com os nós da regra que determinamos na alínea (a).

(c) Viu-se que, por construção, a fórmula que se deduziu na pág. 267 é exacta para o integral $I(f)$, o que não é verdade para a presente fórmula. Por conseguinte, a resposta é negativa. No entanto, são ambas fórmulas de grau máximo, para 3 nós de quadratura em $[-1, 1]$ (levando em consideração as respectivas função peso $w(x)$). A fórmula com nós de Chebyshev usa a função peso $w(x) = \dfrac{1}{\sqrt{1-x^2}}$, enquanto a fórmula deduzida neste exercício usa a função peso $w(x) = 1$. Fórmulas como a que aqui tratamos dizem-se fórmulas de Gauss-Legendre, precisamente por usarem como nós de quadratura os zeros de polinómios de Legendre. ♦

5.8 Leituras recomendadas

M. M. Graça, *A simple derivation of Newton-Cotes formulas with realistic errors*, J. Math. Res., Vol. 4, No. 5, 34-48 (2012).

M. M. Graça, *Quadrature as a least-squares and minimax problem*, Int. J. Numer. Meth. Appl., Vol 10, No. 1, 1-28 (2013). Disponível em http://adsabs.harvard.edu/abs/2012arXiv1206.0281G.

H. Pina, *Métodos Numéricos*, Escolar Editora, 2010, Cap. 4.

Capítulo 6

Equações diferenciais

Um número considerável de problemas importantes em ciência e tecnologia são modelados através de equações diferenciais.

De modo análogo ao que acontece quando pretendemos calcular um determinado integral, também os métodos numéricos para aproximar a solução de uma equação diferencial são imprescindíveis porquanto, em geral, não existem fórmulas explícitas para o seu cálculo, tal como se constatou no Capítulo 5 a respeito do problema de integração numérica.

A área de estudo de métodos numéricos para equações diferenciais é muito vasta. Aqui apenas discutiremos alguns tópicos introdutórios ao tema, pelo que o leitor interessado em aprofundar estas matérias deverá consultar, por exemplo, as obras indicadas na bibliografia.

6.1 Problemas de valor inicial

Vamos iniciar o nosso estudo de métodos numéricos para equações diferenciais ordinárias, de primeira ordem. Relembre-se que uma equação diferencial envolve uma função incógnita y e as suas derivadas. Diz-se equação *ordinária* se a função y é real e de uma única variável real. Uma equação diferencial diz-se de ordem $k \geq 1$ se todas as derivadas que aparecem na equação forem de ordem k ou inferior. Sistemas de equações diferenciais de primeira ordem serão sucintamente tratados na Secção 6.6.

Começamos por equações da forma

$$\begin{array}{l} y'(t) = f(t, y(t)), \qquad t_0 \leq t \leq T \\ y(t_0) = y_0, \end{array} \tag{6.1}$$

onde são dados a função $f : D \subset \mathbb{R}^2 \mapsto \mathbb{R}$, bem como os valores de t_0 e T, e o valor inicial y_0 da solução da equação diferencial. Supõe-se que a solução y é função real definida em $[t_0, T]$, contínua neste intervalo.

Por exemplo, a função contínua que é solução da equação $y'(t) = 2\,y(t)$, tal que $y(0) = -4$, é a função $\phi(t) = -4\,e^{2\,t}$, porquanto $\phi'(t) = 2\,\phi(t)$, e $\phi(0) = -4$. Neste caso, $f(t, y) = 2\,y$, e a equação diferencial diz-nos que a *tangente* à solução y, em cada ponto $(t, y(t))$, possui o valor $2\,y(t)$. Por isso se diz que a função f define um *campo de direcções*.

Assumimos que o domínio D do *campo de direcções* definido pela função f, é o conjunto

$$D = \{(t, y) : \quad t_0 \leq t \leq T, \quad y \in \mathbb{R}\} \subset \mathbb{R}^2, \tag{6.2}$$

ou um subconjunto de D, *convexo*.

As equações (6.1) designam-se habitualmente por *problema de valor inicial* (abreviadamente p.v.i.), visto que do conjunto de soluções possíveis da equação diferencial $y' = f(t, y)$, interessa-nos aquela que satisfaz a *condição inicial* $y(t_0) = y_0$.

O teorema a seguir dá-nos condições suficientes para a existência e unicidade da solução do problema (6.1).

Teorema 6.1. Considere o problema de valor inicial (6.1), onde f está definida no domínio (6.2). Se as funções f e $\partial f/\partial y$ são contínuas em D, então existe pelo menos uma solução.
Se a derivada partial de f em ordem à variável y for limitada em D, isto é, se existir uma constante L, tal que

$$\left|\frac{\partial\, f}{\partial\, y}(t, y)\right| \leq L, \qquad \forall\,(t, y) \in D, \tag{6.3}$$

a solução do p.v.i. é única.

Demonstração. Ver, por exemplo ([5], pág. 142). □

Mesmo quando é possível obter uma fórmula explícita para a solução de um determinado problema do tipo (6.1), isso não significa que fiquemos dispensados de recorrer a métodos numéricos para aproximar os valores dessa solução, tal como acontece no exemplo a seguir.

Exemplo 6.1. *Dado o problema de valor inicial*

$$\begin{array}{l} y'(t) = -e^{t^2}\, y(t), \qquad 1 \leq t \leq 2 \\ y(1) = -1, \end{array} \tag{6.4}$$

(a) Mostrar que existe solução única $y(t)$, e determinar a sua expressão.

(b) Calcular uma aproximação de $y(2)$, aplicando a regra de Simpson com passo $h = 1/4$, usando o integral $\displaystyle\int_1^t e^{s^2}\, ds$.

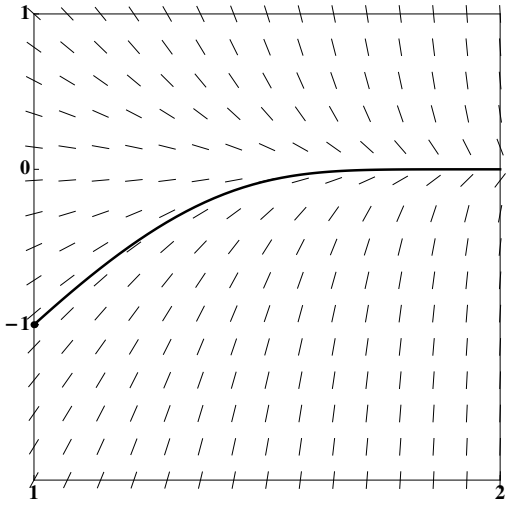


Figura 6.1: Solução do problema de valor inicial (6.4)

(a) Seja $f(t,y) = -e^{t^2}\, y$, onde $1 \leq t \leq 2$, e $y \in \mathbb{R}$. Neste domínio das variáveis t e y, tanto a função f como a sua derivada parcial em ordem a y são funções contínuas. Logo, pelo Teorema 6.1, o p.v.i. dado tem solução contínua no intervalo $[1,2]$. Uma vez que

$$\left| \frac{\partial}{\partial\, y} f(t,y) \right| = e^{t^2} \leq e^4, \qquad \forall t \in [1,2]$$

o mesmo resultado assegura-nos que a solução é única. Na Figura 6.1 está esboçado o campo de direcções da função $f(t,y) = -e^{t^2}\, y$, no domínio $D = [1,2] \times [-2,1]$. A solução do p.v.i. (6.4) está desenhada a traço grosso.

Atendendo a que

$$\frac{y'(t)}{y(t)} = -\, e^{t^2},$$

integrando ambos os membros obtém-se

$$\int_1^t \frac{y'(s)}{y(s)}\, ds = -\int_1^t e^{s^2}\, ds\ .$$

Assim,

$$\int_1^t \frac{y'(s)}{y(s)}\, ds \ \ = \ln\left(y(t)\right) - \ln(-1) = \ln\left(y(t)\right) - \ln\left(e^{i\,\pi}\right)\ .$$

Por conseguinte,

$$y(t) = e^{i\,\pi}\, e^{-\int_1^t e^{s^2}\, ds} = -1\ e^{-\int_1^t e^{s^2}\, ds}\ .$$

Como $y(1) = -1$, a expressão da solução de (6.4) tem a forma

$$y(t) = y(1)\, e^{-\int_1^t e^{s^2}\, ds} = -\, e^{-\int_1^t e^{s^2}\, ds}\ .$$

Não existe uma fórmula explícita para o integral que figura na expressão anterior, pelo que o valor de $y(2)$ terá de ser estimado através de um método numérico.

(b) Seja $F(t) = \int_1^t e^{s^2}\, ds$. Aplicando a regra de Simpson, com $h = 1/4$, temos

$$F(2) \simeq \left[e + e^4 + 4\,(e^{1.25^2} + e^{1.75^2}) + 2\,e^{1.5^2}\right] \simeq 15.0749\ .$$

Assim,

$$y(2) \simeq -\, e^{-15.0749.} \simeq -2.83822 \times 10^{-7}.$$

Recorrendo a uma regra de quadratura mais precisa, pode concluir-se que $y(2) = -3.08984 \times 10^{-7}$ (com 6 algarismos significativos). Por conseguinte, o valor que estimámos para $y(2)$ possui apenas 1 algarismo significativo. Propõe-se ao leitor que determine o passo h que deveria adoptar, caso persistisse em utilizar a regra de Simpson, de modo a garantir, por exemplo, um erro de quadratura inferior a 10^{-13}.

♦

6.2 Método de Euler explícito

Tal como fizemos para as regras de quadratura compostas, comecemos por *discretizar* o problema. O modo mais simples de fazer tal discretização consiste em definir uma "malha" uniforme, que resulta de subdividirmos o intervalo $[t_0, T]$ em N ($N \geq 1$) partes, de comprimento $h = (T - t_0)/N$, considerando os $N + 1$ nós,

$$t_n = t_0 + n\, h, \qquad n = 0 : N.$$

Em cada nó t_n a solução exacta do p.v.i. é $y(t_n)$. Denotamos por y_n um valor aproximado de $y(t_n)$, obtido mediante aplicação de um certo método numérico. Em cada ponto (t_n, y_n), designaremos por *curva integral* a solução do p.v.i. $y' = f(t, y)$, tal que $y(t_n) = y_n$.

O método mais simples para aproximar a solução de (6.1) é o chamado *método de Euler*.[1]

Sabemos que $y'(t_0) = f(t_0, y_0)$, ou seja, que o declive da recta tangente à solução, em t_0, possui o valor $f(t_0, y_0)$. Admitindo que a curva integral que passa em

[1] Leonhard Euler, 1707 – 1783, matemático e físico suiço, considerado um dos cientistas mais eminentes de todos os tempos.

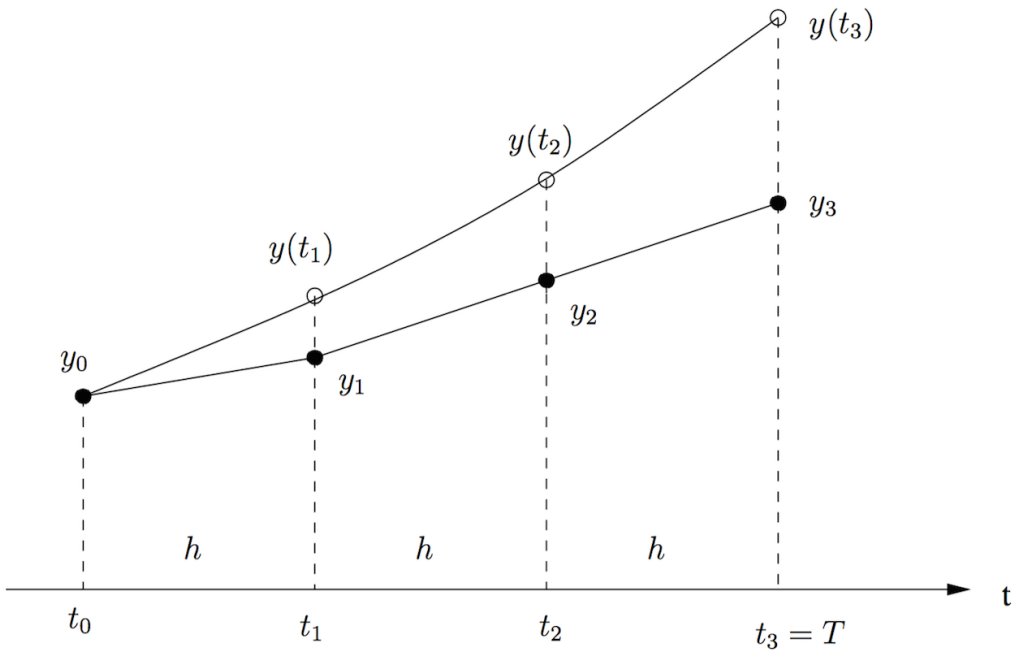


Figura 6.2: Método de Euler com $N = 3$.

(t_0, y_0) é linear entre t_0 e $t_1 = t_0 + h$, aproximemos a solução $y(t)$ por esta linha, seja $\phi_0(t) = y_0 + f(t_0, y_0)\,(t - t_0)$. Em resultado dessa aproximação, o valor exacto $y(t_1)$ é aproximado pelo valor $y_1 = \phi_0(t_1) = y_0 + h\,f(t_0, y_0)$. Por sua vez, a curva integral passando em (t_1, y_1) possui nesse ponto uma tangente cujo declive é $f(t_1, y_1)$. Substituindo essa curva pela sua aproximação linear $\phi_1(t) = y_1 + f(t_1, y_1)\,(t - t_1)$, aproximamos o valor $y(t_2)$ por $\phi_1(t_2)$, ou seja, $y_2 = \phi_1(t_2) = y_1 + h\,f(t_1, y_1)$. O processo é repetido até que seja determinada a aproximação y_{N-1} de $y(t_{N-1})$. No último passo do algoritmo a curva integral passando por (t_{N-1}, y_{N-1}) é aproximada pela função linear $\phi_N(t) = y_{N-1} + f(t_{N-1}, y_{N-1})\,(t - t_{N-1})$ e, finalmente, o valor da solução do p.v.i, em t_N, é aproximado por $y_N = \phi_1(t_{N-1}) = y_{N-1} + h\,f(t_{N-1}, y_{N-1})$.

Em resumo, o método aproxima a solução do problema de valor inicial dado, considerando em cada subintervalo $[t_i, t_i + h]$ a recta tangente à curva integral passando em (t_i, y_i). Assim, o método de Euler é recursivo, da forma

$$\begin{array}{l} y_0 \quad \text{(dado)} \\ y_{n+1} = y_n + h\,f(t_n, y_n), \qquad n = 0 : (N-1)\ . \end{array} \tag{6.5}$$

Os segmentos de recta ligando (t_0, y_0), (t_1, y_1), (t_2, y_2), ..., (t_{N-1}, y_{N-1}) e (t_N, y_N)

definem uma linha "quebrada" como a que se mostra na Figura 6.2, onde o intervalo $[t_0, T]$ foi dividido em $N = 3$ partes.

Caso se considere que o índice n em (6.5) possa ter um qualquer valor inteiro não negativo, a sucessão $(y_n)_{n\geq 0}$ diz-se gerada por uma *equação às diferenças*, de primeira ordem, precisamente porque cada termo da sucessão é definido recursivamente à custa de um só termo anterior.

Aumentando o número N, ou seja, diminuindo o passo h, interessa-nos que as aproximações y_n, definidas pelo esquema (6.5), se aproximem do valor exacto em cada nó, $y(t_i)$, para $i = 1 : N$. Nesse caso, dizemos que o método é *convergente*, segundo a definição a seguir.

Definição 6.1. Um método de aproximação da solução do p.v.i. (6.1) é convergente se e só se

$$\lim_{h\to 0} |y(t_i) - y_i| = 0, \qquad 0 \leq i \leq N \ . \tag{6.6}$$

O método (6.5) é *explícito*, porquanto o valor novo y_{n+1} depende explicitamente do anterior y_n. Já o método a que a seguir faremos referência, define o valor y_{n+1} de modo implícito e, por isso, se diz *método de Euler implícito*.

Método de Euler implícito

Equações às diferenças para aproximar a solução do p.v.i. (6.1) podem ser obtidas recorrendo a regras de quadratura. Em particular, as regras de quadratura mais simples, como a *regra do rectângulo à esquerda* e a *regra do rectângulo à direita*, referidas na pág. 276, permitem-nos deduzir facilmente as expressões respectivamente do método de Euler explícito e implícito.

Com efeito, de (6.1), resulta imediatamente

$$\int_{t_n}^{t_{n+1}} y'(s)\, ds = \int_{t_n}^{t_{n+1}} f(s, y(s))\, ds \ . \tag{6.7}$$

Se na expressão anterior aproximarmos o integral à direita usando a regra do rectângulo à esquerda, admitindo que $f(t_n, y(t_n)) = y_n$, obtém-se

$$y_{n+1} = y_n + h\, f(t_n, y_n),$$

ou seja, a equação às diferenças do método de Euler explícito.

Se em (6.7) aplicarmos a regra do rectângulo à direita, supondo que $y_{n+1} = f(t_{n+1}, y(t_{n+1}))$, obtém-se o método de Euler implícito,

$$\begin{array}{ll} y_0 & \text{(dado)} \\ y_{n+1} & = y_n + h\, f(t_{n+1}, y_{n+1}) \qquad n = 0 : (N-1) \ . \end{array} \tag{6.8}$$

Os métodos implícitos para problemas de valor inicial, como é o caso do método (6.8), levam-nos a relembrar o que estudámos a respeito de métodos do ponto fixo. Com efeito, as equações

$$y_{n+1} = g(y_{n+1}) = y_n + h\, f(t_{n+1}, y_{n+1})$$

são equações de ponto fixo, com incógnita y_{n+1}. De facto, em cada passo do método de Euler implícito devemos resolver uma equação de ponto fixo

$$y = g(y) = \alpha + h\, f(t_{n+1}, y), \quad \text{com} \quad \alpha = y_n, \qquad y \in \mathbb{R}\ . \tag{6.9}$$

Sendo válidas as hipóteses do Teorema 6.1, temos

$$|g'(y)| \leq h\, L\ . \tag{6.10}$$

Assim, escolhendo um passo $h < \dfrac{1}{L}$, o método de ponto fixo gerado pela função iteradora em (6.9) é localmente convergente para um ponto fixo atractor (ou excepcionalmente superatractor).

Dado que uma solução y da equação (6.9) é uma aproximação da solução do p.v.i. no ponto $t = t_{n+1}$, é usual considerar-se para estimativa inicial do processo iterativo de ponto fixo,

$$y^{(k+1)} = g(y^{(k)}), \qquad k = 0, 1, \ldots, \tag{6.11}$$

(ou seja, um valor inicial "suficientemente próximo" do ponto fixo), o valor $y^{(0)} = y_n$, sendo y_n obtido mediante um passo do método de Euler explícito com início em y_{n-1}, e efectuar algumas iterações do processo (6.11), tal como se ilustra no Exemplo 6.2, p. 294.

6.2.1 Erro do método de Euler explícito

Uma vez satisfeitas as hipóteses do Teorema 6.1, pág. 286, sabemos que o p.v.i. (6.1) possui solução única. Além disso, se forem conhecidas constantes L e M, limitando respectivamente a derivada partial $\partial f/\partial y$ e a função y'', podemos garantir convergência do método de Euler, conforme se prova adiante.

Desprezando erros de arredondamento, a fórmula de majoração de *erro global* que é deduzida na demonstração do Teorema 6.2 a seguir, habilita-nos a prever uma escolha do passo h do método de Euler, de modo a garantir que o erro absoluto das aproximações da solução $y(t)$, calculadas em todo o intervalo $[t_0, T]$, sejam não superiores a uma tolerância prefixada.

Teorema 6.2. Seja $h > 0$ o passo do método de Euler (6.5) aplicado ao problema de valor inicial (6.1), de modo que num domínio convexo $D \subset \mathbb{R}^2$ sejam satisfeitas as desigualdades

$$\max_{\substack{t_0 \le t \le t_i \\ y \in \mathbb{R}}} \left| \frac{\partial f}{\partial y}(t,y) \right| = L, \quad \text{e} \quad \max_{t_0 \le t \le T} |y''(t)| = M, \tag{6.12}$$

assumindo que y'' é suficientemente regular, no sentido de que $y''(t) \in C^2([t_0, T])$. Desprezando erros de arredondamento, em cada ponto $t_n = t_0 + n\ h$, da malha definida no intervalo $[t_0, T]$, o erro absoluto da aproximação y_n satisfaz a desigualdade

$$|e_n| = |y(t_n) - y_n| \le \frac{M}{2\,L} \left(e^{L\,(t_n - t_0)} - 1 \right)\ h\ . \tag{6.13}$$

Consequentemente, o método de Euler converge, existindo uma constante $C > 0$, tal que

$$||e_h||_\infty = \max_{0 \le n \le N} |y(t_n) - y_n| \le C\,h, \quad \text{isto é,} \quad ||e_h||_\infty = \mathcal{O}(h)\ . \tag{6.14}$$

Demonstração. Seja $t = t_n$ um qualquer ponto da malha uniforme considerada. Considere-se o desenvolvimento de Taylor de primeira ordem, em torno de t_n, da solução $y(t)$. Podemos escrever,

$$y(t_{n+1}) = y(t_n) + h\,f(t_n, y(t_n)) + \frac{h^2}{2}\,y''(\xi_n), \qquad \xi_n \in (t_n, t_{n+1})\ . \tag{6.15}$$

Subtraindo membro a membro com a equação às diferenças do método de Euler,

$$y_{n+1} = y_n + h\,f(t_n, y_n),$$

resulta

$$y(t_{n+1}) - y_{n+1} = y(t_n) - y_n + h\,[f(t_n, y(t_n)) - f(t_n, y_n)] + \frac{h^2}{2}\,y''(\xi_n)\ . \tag{6.16}$$

Como por hipótese f e $\partial f/\partial y$ são funções contínuas no domínio convexo D, podemos aplicar o teorema de Lagrange tomando y como variável independente, e assim garantir a existência de pelo menos um ponto $\eta_n \in int(y(t_n), y_n)$, tal que

$$f(t_n, y(t_n)) - f(t_n, y_n) = \frac{\partial f}{\partial y}(t_n, \eta_n) \times (y(t_n) - y_n)\ .$$

Por conseguinte, a igualdade (6.16), permite-nos comparar os erros e_{n+1} e e_n,

$$\begin{aligned} e_{n+1} &= e_n + h\,\frac{\partial f}{\partial y}(t_n, \eta_n)\,e_n + \frac{h^2}{2}\,y''(\xi_n) \\ &= \left(1 + h\,\frac{\partial f}{\partial y}(t_n, \eta_n)\right)\ e_n + \frac{h^2}{2}\,y''(\xi_n)\ . \end{aligned} \tag{6.17}$$

Considerando erros absolutos, e entrando em consideração com as majorações (6.12) de $|\partial f/\partial y|$ e de $|y''|$, obtém-se,

$$\begin{array}{ll} |e_0| & = |y(t_0) - y_0| = 0 \\ |e_{n+1}| & \leq (1 + h\,L)\;|e_n| + \dfrac{M}{2}\,h^2, \qquad n = 0 : (N-1)\;. \end{array} \tag{6.18}$$

Sejam

$$a = 1 + h\,L \geq 1, \qquad b = \frac{M}{2}\,h^2 \geq 0$$

As desigualdades (6.18) são da forma

$$|e_{n+1}| \leq a\,|e_n| + b, \qquad n = 0 : (N-1)\;.$$

Assim,

$$\begin{array}{ll} |e_1| & \leq b \\ |e_2| & \leq a\,b + b = (a+1)\,b \\ |e_3| & \leq a^2\,b + a\,b + b = (a^2 + a + 1)\,b \\ & \vdots \\ |e_k| & \leq (a^{k-1} + a^{k-2} + \ldots + a + 1)\,b, \qquad k = 1 : N\;. \end{array}$$

No segundo membro da desigualdade anterior encontra-se entre parêntesis uma soma geométrica de razão a. Por conseguinte,

$$|e_k| \;\; \leq \frac{a^k - 1}{a - 1} \times b = \frac{(1 + h\,L)^k - 1}{h\,L} \times \frac{M}{2}\,h^2,$$

ou seja,

$$|e_k| \leq \frac{M}{2\,L}\;\left[(1 + h\,L)^k - 1\right]\;h\;. \tag{6.19}$$

O desenvolvimento de Taylor da função exponencial permite-nos escrever a soma

$$e^{h\,L} = 1 + h\,L + \frac{(h\,L)^2}{2!} + \frac{(h\,L)^3}{3!} + \ldots.$$

Logo,

$$1 + h\,L < e^{h\,L} \quad \Longrightarrow \quad (1 + h\,L)^k < e^{k\,h\,L}.$$

Substituindo a última desigualdade em (6.19), obtém-se

$$|e_k| \leq \frac{M}{2\,L}\;\left(e^{k\,h\,L} - 1\right)\;h, \qquad k = 1 : N\;.$$

Dado que $t_k - t_0 = k\,h$, resulta a majoração de erro absoluto em t_k,

$$|e_k| \leq \frac{M}{2\,L}\;\left(e^{L(t_k - t_0)} - 1\right)\;h,$$

h	Aprox. em 2.0	Erro
0.2	-3.14164	0.00559
0.1	-3.14019	0.00414
0.05	-3.13829	0.00224
0.025	-3.13720	0.00115

Tabela 6.1: Método de Euler explícito – Exemplo 6.2.

e, no intervalo $[t_0, T]$,

$$|e_k| \leq \frac{M}{2\,L}\,\left(e^{L(T-t_0)}-1\right)\,h, \qquad k=1:N\ .$$

Das desigualdades anteriores conclui-se que $\lim_{h\to 0}|e_k| = 0$, para $k = 0 : N$, ou seja, o método converge. Fazendo $C = \dfrac{M}{2\,L}\,\left(e^{L(T-t_0)}-1\right)$, fica mostrada a validade das relações (6.16). □

No Exemplo 6.2 a seguir, é efectuada uma aplicação do Teorema 6.2. A determinação das constantes L e M em (6.13) é por vezes laboriosa e, frequentemente, leva-nos a determinar majorações de erro de truncatura manifestamente excessivas. Por conseguinte, o referido teorema tem sobretudo interesse teórico porquanto nos dá condições suficientes para a convergência do método de Euler explícito.

Exemplo 6.2. *Considere-se o problema de valor inicial,*

$$y'(t) = e^t\,\sin(y(t)), \qquad y(-2) = -1/2,$$

com $t \in [-2, 2]$.
(a) Efectuando um passo do método de Euler explícito, calcular uma aproximação y_1 *da solução do p.v.i. dado, no ponto* $t = -1.8$. *Repetir utilizando o método de Euler implícito.*

(b) Recorrendo à expressão (6.13)*, obter uma majoração do erro do valor* y_1*, calculado na alínea anterior pelo método de Euler explícito.*

(c) Sabendo que o valor exacto da solução é $y(-1.8) = -0.514555$ *(6 algarismos significativos), concluir qual das aproximações calculadas na alínea (a) é mais precisa.*

(d) No intervalo $[t_0, T] = [-2, 2]$*, pretende-se obter gráficos (análogos aos da Figura 6.3, para a malha uniforme* $(t_i, y_i)_{i=0}^{i=N}$ *que resulta da aplicação do método de Euler explícito, respectivamente com passo* $h = 0.2/2^j$, $j = 0 : 3$. *Verificar os resultados da Tabela 6.1, onde se dão as aproximações obtidas pelo método, no ponto* $t = 2.0$ *(valores arredondados para 6 dígitos, sabendo que* $y(2) = -3.13605$*).*

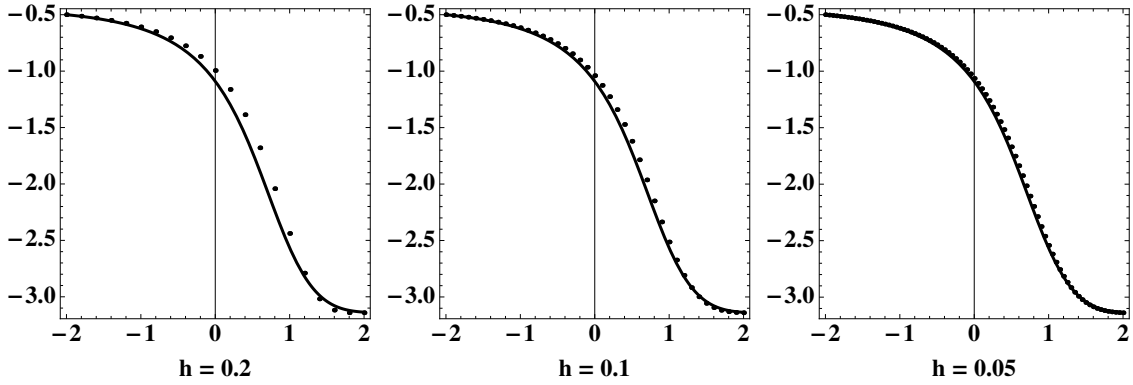


Figura 6.3: Convergência do método de Euler (Exemplo 6.2). O gráfico da solução $y(t)$ está desenhado a traço cheio. Os pontos representam aproximações obtidas através do método de Euler explícito.

É ou não verdade que os valores tabelados, bem como os gráficos da Figura 6.3, sugerem convergência do método de Euler no intervalo considerado?

(a) A função definindo o campo de direcções associado ao p.v.i. proposto é $f(t, y) = -e^t \sin(y)$. Por conseguinte, a equação às diferenças correspondente ao método de Euler explícito escreve-se,

$$\begin{aligned} &y_0 = -1/2 \\ &y_{n+1} = y_n + h\, e^{t_n} \sin(y_n), \qquad n = 0, 1, \ldots \end{aligned}$$

Para $h = 0.2$ e $t_0 = -2$, obtém-se

$$y(-1.8) \simeq y_1 = -0.5 + 0.2\, e^{-2} \times \sin(-0.5) = -0.512977 \ .$$

O método de Euler implícito (6.8), pág. 290, tem a forma

$$y_{n+1} = y_n + h\, e^{t_{n+1}} \sin(y_{n+1}) \ .$$

Para $y_0 = -1/2$ e $t_1 = -2 + h = -1.8$, a aproximação y_1 deste método é solução do problema de ponto fixo,

$$y = g(y) = -0.5 + 0.2\, e^{-1.8} \sin(y) \ \Rightarrow g'(y) \simeq 0.0330598 \cos(y), \quad \forall y \in \mathbb{R} \ .$$

Como $0 < g'(y) << 1$, o método de ponto fixo terá convergência linear, rápida, e monótona. Com efeito, aproveitando a aproximação de $y(-1.8)$ calculada anteriormente pelo método de Euler explícito, ou seja tomando para aproximação inicial $y^{(0)} = -0.512977$, são as seguintes as primeiras 2 iteradas do método de ponto fixo $y^{(k+1)} = g(y^{(k)})$:

$$\begin{aligned} y^{(0)} &= -0.512977 \\ y^{(1)} &= -0.516225 \\ y^{(2)} &= -0.516318 \ . \end{aligned}$$

Tomemos para estimativa da solução do p.v.i. pelo método de Euler implícito, em $t = -1.8$, o último valor da lista de iteradas anteriores, ou seja, $y_1 = -0.516318$. O erro da última iterada do método de ponto fixo, relativamente à respectiva solução, é

$$|y - y_1| \leq |y^{(2)} - y^{(1)}| < 10^{-3}.$$

Visto que $y(-1.8) = -0.514555$, o erro anterior é muito menor do que o erro de truncatura $|e_1| = |y(-1.8) - y_1|$, pelo que as duas iterações que efectuámos do método de ponto fixo são suficientes para o fim em vista.

(b) As majorações de erro do método de Euler obtidas a partir da expressão (6.13), pág. 292, possuem o inconveniente de serem frequentemente difíceis de obter (nomeadamente o cálculo da constante M) e/ou levam-nos a estimativas de erro por vezes demasiado grosseiras no intervalo $[t_0, T]$. No presente caso, restringimos o intervalo a $[t_0, T] = [-2, -1.8]$. Convida-se o leitor a calcular uma estimativa do erro global no intervalo $[-2, 2]$.

Dado que $\partial f/\partial y = e^t \cos(y)$, no intervalo $[-2, 2]$, tem-se

$$L = \max \left| \frac{\partial f}{\partial y}(t, y) \right| \leq e^2, \qquad \forall\, y \in \mathbb{R} \ .$$

A partir da expressão de y', obtém-se

$$y''(t) = e^t \, \sin(y(t)) \, \left(1 + e^t \, \cos(y(t))\right) ,$$

donde,

$$M = max_{-2 \leq t \leq 2} \, |y''(t)| \leq e^2(1 + e^2) \ .$$

Aplicando a desigualdade (6.13), para $t = t_1$ e $h = 0.2$, obtém-se,

$$\begin{aligned} |e_1| = |y(-1.8) - y_1| \ &\leq \frac{M}{2\,L} \left(e^{L \times 0.2} - 1\right) \times 0.2 \\ &\leq \frac{1 + e^2}{2} \, \left(e^{e^2 \times 0.2} - 1\right) \times 0.2 \simeq 2.84 \ . \end{aligned}$$

O valor anteriormente calculado é desprovido de interesse prático porquanto o erro de truncatura cometido é, de facto, muito inferior, conforme se mostra a seguir.

(c) O erro no ponto $t = -1.8$, com passo $h = 0.2$, para o método de Euler explícito é

$$|y(-1.8) - y_1| = | - 0.514555 + 0.512977| \simeq 0.0016,$$

e para o método implícito,

$$|y(-1.8) - y_1| = | - 0.514555 + 0.516318| \simeq 0.0018 \ .$$

Assim, neste caso, o método de Euler explícito produz um resultado mais preciso.

(d) A expressão $||e_h||_\infty = \mathcal{O}(h)$ em (6.16), diz-nos que, para h suficientemente pequeno, o erro global no método de Euler é aproximadamente reduzido a metade, se em vez do passo h usarmos, por exemplo, o passo $h/2$. Um método convergente que possua este tipo de comportamento diz-se um método de *primeira ordem* de convergência, segunda a Definição 6.2 dada adiante, pág. 298. A última coluna da Tabela 6.1 mostra que o erro calculado no ponto $t = 2.0$ é, aproximadamente, reduzido a metade quando passamos de $h = 0.05$ a $h = 0.025$, confirmando ser 1 a ordem de convergência do método de Euler aplicado ao problema em causa.

♦

Erro local do método de Euler

Admitamos que a solução y do problema de valor inicial (6.1) (pág. 285) é, pelo menos, de classe $C^2([t_0, T])$. Fixado um nó t_n em $[t_0, T]$, compare-se o valor exacto $y(t_{n+1})$, após um passo do método de Euler, com o valor calculado y_{n+1}. É válido o desenvolvimento de Taylor,

$$y(t_{n+1}) = y(t_n) + h\, f(t_n, y(t_n)) + \frac{h^2}{2}\, y''(\xi_n), \qquad \xi_n \in (t_n, t_{n+1}) \ .$$

Supondo que $y_n = f(t_n, y(t_n))$ – ou seja, que o passo do método tem início no ponto exacto $(t_n, y(t_n))$) – o erro, T_{n+1}, cometido neste passo, é

$$T_{n+1} = y(t_{n+1}) - y_{n+1} = \frac{h^2}{2}\, y''(\xi_n), \qquad \xi_n \in (t_n, t_{n+1}) \ .$$

Considerando o erro local absoluto, e fazendo $M = \max_{t_0 \le t \le T} |y''(t)|$, obtém-se a majoração

$$|T_{n+1}| = |y(t_{n+1}) - y_{n+1}| \le \frac{M}{2}\, h^2, \qquad n = 0 : (N-1) \ . \tag{6.20}$$

Assim, para h suficientemente pequeno, o erro local em cada ponto da malha é da ordem de h^2 (enquanto que o erro global é da ordem de h^1, como vimos em (6.16), pág. 292).

Ordem de convergência

A expressão (6.16), pág. 292, indica que o erro global do método de Euler é proporcional a h^1, e por isso se diz que este método possui ordem de convergência um, de acordo com a Definição a seguir.

Definição 6.2. Um método numérico convergente para a solução do problema de valor inicial (6.1) diz-se possuir ordem de convergência $p > 0$ se, para um passo h suficientemente pequeno, existir uma constante $C > 0$ tal que

$$||e_h||_\infty = \max_{0 \le i \le N} |y(t_i) - y_i| \le C\, h^p,$$

onde $N = (T - t_0)/h$. A constante C é independente de h, embora possa depender de f e do intervalo $[t_0, T]$ considerado.

6.3 Métodos de Taylor

O método de Euler (6.5) é um caso particular de métodos de ordem de convergência $p \ge 1$, designados por *métodos de Taylor*.

Supondo que a solução $y(t)$ do p.v.i. (6.1) é suficientemente regular, o método de Euler pode obter-se retendo os dois primeiros termos do desenvolvimento de Taylor de ordem $p \ge 1$. Sendo $h = t_{n+1} - t_n$, tem-se

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + h\, y'(t_n) + \frac{h}{2}\, y''(t_n) + \dots \\ &= y(t_n) + h\, f(t_n, y(t_n)) + \frac{h^2}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\, f\right](t_n, y(t_n)) + \dots + \\ &+ \frac{h^p}{p!} y^{(p)}(t_n) + \frac{h^{p+1}}{(p+1)!}\, y^{(p+1)}(\xi), \quad \xi \in (t_n, t_{n+1}) \ . \end{aligned} \tag{6.21}$$

Fazendo $y_n = y(t_n)$, e desprezando o resto do desenvolvimento (6.21), obtém-se a equação às diferenças do método de Taylor de ordem p,

$$y_{n+1} = y_n + h\, f(t_n, y_n) + \dots + \frac{h^p}{p!}\, f^{(k-1)}(t_n, y_n) \ . \tag{6.22}$$

Para $p = 1$ resulta o método de Euler explícito.

Fixado um passo suficientemente pequeno $h > 0$, mostra-se que o erro local do método (6.22) é proporcional a h^{p+1}, enquanto que o respectivo erro global é proporcional a h^p. Tal significa que, caso o método (6.22) convirja, trata-se de um método de ordem de convergência p, segundo a Definição 6.2.

No Exemplo 6.3 a seguir, compara-se o método de Euler com o método de Taylor de segunda ordem. A baixa precisão dos resultados obtidos pelo método de Euler, explica por que razão este método é geralmente preterido a favor de métodos de ordem de convergência superior.

Os métodos de Taylor de ordem $p \ge 2$, no entanto, possuem o inconveniente de necessitarem do cálculo das sucessivas derivadas parciais, implícitas no símbolo

$y^{(p)} = f^{(p-1)}$ nas expressões (6.21) e (6.22), pelo que métodos dispensando derivação parcial da funcão $f(t, y)$ são em geral preferíveis. Tais métodos serão sucintamente abordados na Secção 6.4, pág. 304.

Exemplo 6.3. *O problema de valor inicial*

$$\begin{aligned} &y(0) = 1/2 \\ &y'(t) = 1 + (y(t) - t)^2, \qquad 0 \le t \le 1, \end{aligned}$$

tem solução

$$y(t) = \frac{t^2 - 2\,t - 1}{t - 2}\ .$$

(a) Obter um valor aproximado de $y(0.3)$*, aplicando respectivamente o método de Euler e de Taylor de segunda ordem, com passo* $h = 0.1$*.*

(b) Utilizando a função Sig, definida em (3.194)*, pág. 179, comparar graficamente o número de algarismos significativos dos valores calculados pelos métodos referidos, numa malha de passo* $h = 1/10$*,* $h = 1/20$ *e* $h = 1/40$*.*

(a) A função $f(t, y) = 1 + (y - t)^2$ é regular para $t \in [0, 1]$, e $y \in \mathbb{R}$. Tem-se

$$\begin{aligned} y''(t) = f^{(1)}(t, y) \ &= \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\, f\right)(t, y) \\ &= \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y)\, f(t, y) \\ &= -2\,(y - t) + 2\,(y - t)\ [1 + (y - t)^2] = 2\,(y - t)^3. \end{aligned}$$

Assim, o método de Euler escreve-se

$$y_{n+1} = y_n + h\,(1 + (y_n - t_n)^2)\ .$$

O método de Taylor de segunda ordem "corrige"o método anterior, obtendo-se

$$y_{n+1} = y_n + h\,(1 + (y_n - t_n)^2) + h^2\,(y_n - t_n)^3,$$

onde $y_0 = 1/2$ e $h = 0.1$.

A Tabela 6.2 mostra os valores calculados e respectivos erros para o método de Euler (à esquerda) e o método de Taylor de ordem 2 (à direita). O erro global em $t = 0.3$ deste último método é cerca de 10 vezes menor do que o correspondente erro do método de Euler.

(b) Na Figura. 6.4 compara-se o número de algarismos significativos dos valores calculados para os dois métodos, no intervalo $[0, 1]$, e para o passo h indicado.

É evidente a grande vantagem do método de Taylor de segunda ordem relativamente ao método de Euler. Note-se que no ponto $t = 1.0$, para o passo $h = 1/40$,

t_i	y_i	$\lvert y(t_i) - y_i \rvert$	y_i	$\lvert y(t_i) - y_i \rvert$
0.1	0.6250000	0.0013158	0.6262500	0.00006579
0.2	0.7525625	0.0029931	0.75540130	0.0001542
0.3	0.8830950	0.00514026	0.8879616	0.00027370

Tabela 6.2: Comparação do método de Euler (colunas da esquerda) com o método de Taylor de segunda ordem (colunas da direita) – ver Exemplo 6.3.

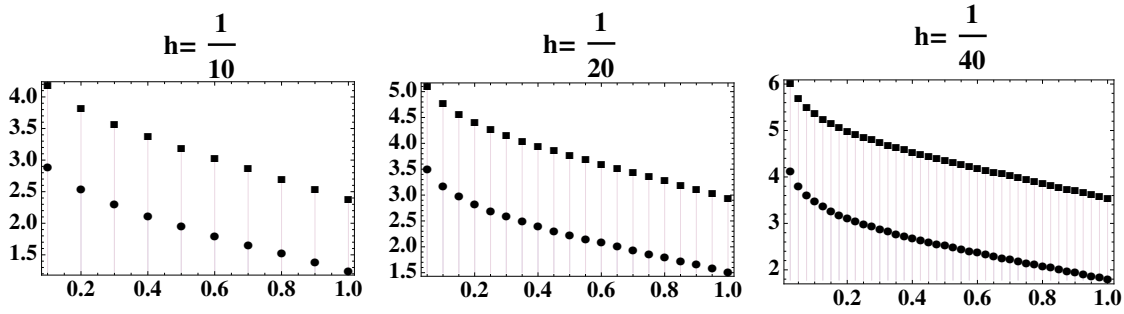


Figura 6.4: Número de algarismos significativos aproximados dos métodos de Euler (valores assinalados com quadrados negros a cheio) e Taylor de segunda ordem (valores assinalados com círculos a cheio) – ver Exemplo 6.3)

o método de Euler produz um resultado com menos do que 2 algarismos significativos (aproximadamente). A baixa precisão deste método evidenciada neste exemplo, explica a razão pela qual na prática se preferem métodos de ordem de convergência superior a um. Entre estes, avultam os chamados métodos de Runge-Kutta que serão discutidos na secção 6.4.

♦

6.3.1 Simulação do erro global

Na prova do Teorema 6.2, pág. 292, foram deduzidas majorações do erro global para o método de Euler explícito. Tal como se ilustrou através do Exemplo 6.2, tais majorações são geralmente laboriosas senão impossíveis de realizar e, frequentemente, pouco realistas dado sobreavaliarem o erro realmente produzido.

Adoptando a situação que usualmente ocorre na prática, ou seja, quando a solução de um determinado p.v.i. não é conhecida, é possível simular o erro global de um método recorrendo eventualmente a computação simbólica. Para tanto, iremos simular o erro global de um determinado método de ordem $p \geq 1$ de convergência, aproximando convenientemente a respectiva equação às diferenças que modele teoricamente o respectivo erro.

Por exemplo, para o método de Euler explícito, vamos usar a equação às diferenças (6.17), a qual modela o erro global deste método. O objectivo é aproximar esse modelo teórico de modo a determinar estimativas *realistas* do erro $e_k = y(y_k) - y_k$, no intervalo $[t_0, T]$, onde pretendemos determinar a solução $y(t)$ de um problema de valor inicial.

Definição 6.3. (Estimativa realista de erro)
Dizemos que uma estimativa do erro $e_k = y(t_k) - y_k$, produzida por um método de ordem $p \geq 1$ de convergência é *realista*, se o modelo de equação às diferenças utilizado para calcular essa estimativa produzir um erro estimado aproximadamente igual a $h/2^p$, quando no método em causa passamos do passo h ao passo $h/2$, para h suficientemente pequeno.

Ao determinarmos aproximações y_k da solução (desconhecida) $y(t_k)$, mediante um processo numérico de ordem p, faremos acompanhar os valores calculados de y_k pela respectiva estimativa realista de erro. Se o método for convergente, os erros realistas estimados simularão bem os erros associados à equação às diferenças que modelam o erro teórico do método usado. A análise do erro simulado no intervalo $[t_0, T]$ irá permitir inferir a convergência do método em causa e confirmar a respectiva ordem de convergência.

No Exemplo 6.4 adiante, retomamos o p.v.i. tratado no Exemplo 6.2, pág. 294, para o qual sabemos não se conhecer a expressão da respectiva solução $y(t)$,

no intervalo $[-2, 2]$, obtendo erros realistas para o método de Euler explícito, mediante aplicação do modelo de erro aproximado que a seguir se descreve.

O processo pode ser generalizado a métodos de ordem de convergência superior a um, desde que se conheça o respectivo modelo teórico para o erro global.

Boas estimativas do erro global de um método numérico para equações diferenciais são indispensáveis em tecnologia espacial. Por exemplo, deve-se ao astrónomo e matemático argentino Pedro Zaduraisky [39] um método geral para cálculo aproximado do referido erro. As técnicas numéricas de Zaduraisky permitiram-lhe, nomeadamente, calcular a órbita do primeiro satélite artificial americano *Explorer I*, bem como a órbita de uma das luas de Saturno.

Estimativas realistas do erro global do método de Euler

Os símbolos f_1' e f_2' usados a seguir designam repectivamente derivação parcial em ordem à primeira e segunda variáveis. A partir da expressão (6.17), pág. 292, substituindo o ponto desconhecido η_n por y_n, e o ponto desconhecido ξ_n por t_n, resulta imediatamente a equação às diferenças,

$$\begin{aligned} &e_0 = 0 \\ &e_{n+1} = (1 + h\, f_2'(t_n, y_n))\; e_n + \frac{h^2}{2}\, y''(t_n), \quad n = 0, 1, \dots \ . \end{aligned} \tag{6.23}$$

A equação às diferenças anterior aproxima a equação às diferenças teórica que modela o erro do método em causa.
Atendendo a que

$$y''(t_n) = (f_1' + f_2'\, f)\,(t_n, y_n),$$

a equação às diferenças que nos servirá de modelo para o cálculo de erros realistas do método de Euler explícito, tem a forma

$$\begin{aligned} &e_0 = 0 \\ &e_{n+1} \simeq (1 + h\, f_2'(t_n, y_n))\; e_n + \frac{h^2}{2}\; (f_1'(t_n, y_n) + f_2'(t_n, y_n)\, f(t_n, y_n))\,, \quad n = 0, 1, \dots \ . \end{aligned} \tag{6.24}$$

Uma vez decidido experimentalmente se a equação às diferenças aproximada (6.24) produz ou não estimativas realistas para o erro do método de Euler aplicado a um problema concreto, isto é, caso se verifique experimentalmente que o erro global é aproximadamente reduzido a metade quando passamos de um determinado passo h ao passo $h/2$, podemos concluir que o modelo de erro (6.24) simula correctamente o modelo teórico de erro (6.17).

Note-se que se substituirmos os valores y_k calculados pelo método de Euler, pelos valores

$$\tilde{y}_k = y_k + e_k, \tag{6.25}$$

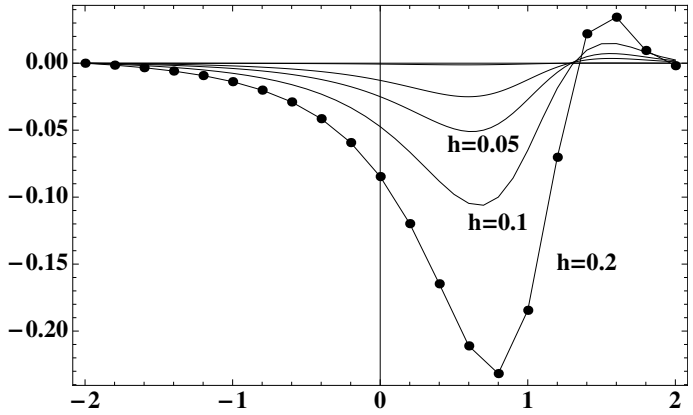


Figura 6.5: Estimativas realistas de erro para o método de Euler (ver Exemplo 6.4).

h	y_k	Erro realista	Erro exacto
0.2	−2.04209	−0.231728	−0.197483
0.1	−2.14944	−0.100010	−0.0901402
0.05	−2.19672	−0.043599	−0.0428571
0.025	−2.21869	−0.0215077	−0.020884

Tabela 6.3: Comparação do erro realista com o erro exacto para o método de Euler, em $t = 0.8$ – ver Exemplo 6.4.

onde e_k é uma estimativa realista de erro calculada a partir de (6.24), o valor $\tilde{y}_k$ é o mesmo que obteria caso tivesse aplicado o método de Taylor de ordem 2 ao p.v.i em causa.

Exemplo 6.4. *Considerando o p.v.i. do Exemplo 6.2, pág. 294, aplicar o método de Euler explícito para os valores do passo* $h = 0.2$*,* $h = 0.1$*,* $h = 0.05$ *e* $h = 0.025$*. Concluir graficamente que as respectivas estimativas de erro* (6.24) *são realistas no intervalo* $[-2, 2]$*.*

Na Figura. 6.5 apresentam-se as curvas ligando os pontos (t_k, e_k), onde o erro realista e_k foi calculado recursivamente aplicando a fórmula (6.24), para cada um dos valores de h indicados. O gráfico obtido não só sugere a convergência do método (o erro global tende para o valor nulo à medida que o passo diminui), como nos mostra que o modelo de erro aproximado (6.24) simula bem o facto do método de Euler ser de primeira ordem de convergência para o p.v.i. em causa. Com efeito, o erro calculado, por exemplo nas imediações do ponto $t = 0.8$, onde esse erro tem maior grandeza, passa aproximadamente a metade do seu valor quando passamos de $h = 0.1$ para $h = 0.05$, confirmando o que já se tinha observado a respeito do erro exacto (ver pág. 295).

Na Tabela 6.3 compara-se o erro realista com o erro exacto, calculado em $t = 0.8$ para cada um dos valores de h considerados. Além de nos dar o sinal correcto, em toda a gama de valores de h usados, o erro realista possui pelo menos um algarismo significativo por comparação com o erro exacto. Assim, usando os valores calculados para $h = 0.025$, tem-se

$$y(0.8) \simeq -2.21869 - 0.020884 = -2.23957.$$

O valor anterior possui pelo menos 3 algarismos significativos (na realidade possui 5, porquanto o valor exacto arredondado para 6 dígitos é $y(0.8) = -2.23958$).

Este exemplo mostra-nos que o cálculo dos valores y_k do método de Euler, acompanhados dos respectivos erros realistas, pode revelar muito acerca de potenciais dificuldades de natureza numérica inerentes ao problema de valor inicial proposto. Caso o erro estimado tenha o comportamento próprio do método utilizado (neste caso, um método de primeira ordem) tal significa que a solução do problema é "bem comportada", enquanto que um erro estimado em desacordo com o que a teoria faz prever, pode querer significar a ocorrência de uma solução que não satisfaz os pressupostos do Teorema 6.2 no intervalo $[t_0, T]$, ou seja, para a qual o modelo de erro exacto (6.17), pág. 292, não é válido. ⧫

6.4 Métodos de Runge-Kutta de segunda ordem

A fim de obtermos uma expressão para uma família de métodos de segunda ordem de convergência, capazes de aproximar a solução de um problema de valor inicial, descreve-se a seguir uma certa combinação linear de funções. O objectivo é substituir o método de Taylor de ordem dois, referido no parágrafo 6.3, pág. 298, por um método aproximado, também de segunda ordem, mas que não utilize derivação parcial. A família de métodos desse tipo, discutida a seguir, recebe a designação de métodos de Runge-Kutta de segunda ordem.

Supondo que a solução y do p.v.i. considerado é tal que $y \in C^3([t_0, T])$, e $h > 0$ é um dado passo, relembre-se que o método de Taylor de segunda ordem se escreve,

$$\begin{aligned} y(t+h) &= y(t) + h\, f(t, y(t)) + \frac{h^2}{2}\, (f_1'(t, y(t)) + f_2'(t, y(t))\, f(t, y(t))) + \mathcal{O}(h^3) \\ &= y(t) + h\, F(t, y) + \mathcal{O}(h^3), \end{aligned} \tag{6.26}$$

onde

$$F(t, y) = f(t, y(t)) + \frac{h}{2}\, (\, f_1'(t, y(t)) + f_2'(t, y(t))\, f(t, y(t))\,)\,. \tag{6.27}$$

Pretende-se aproximar a função $F(t, y)$, por outra $\bar{F}(t, y)$, de modo que o respectivo erro de truncatura seja da ordem $\mathcal{O}(h^2)$. Como em (6.26), a expressão de

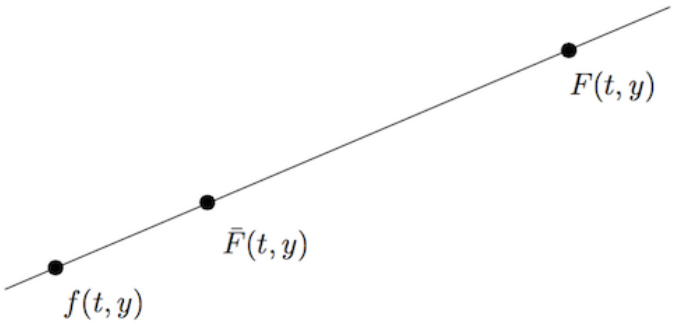


Figura 6.6: Combinação linear de $f(t,y)$ e $\bar{F}(t,y)$.

$F(t,y)$ aparece multiplicada por h, o erro final será da mesma ordem de grandeza do erro de truncatura em (6.26), ou seja, $\mathcal{O}(h^3)$.

Seja $\alpha \neq 0$ um parâmetro a determinar, e considere-se como modelo a função $\bar{F}$, tal que

$$\bar{F}(t,y) = f(t + \alpha\, h, y + \alpha\, h f(t,y)), \tag{6.28}$$

a qual, enquanto função de α, possui como desenvolvimento de Taylor de segunda ordem, em torno de $\alpha = 0$,

$$\bar{F}(t,y) = f(t,y) + \alpha\, h\, f_1'(t,y) + \alpha\, h\, f_2'(t,y) + \mathcal{O}((\alpha\, h)^2)\ . \tag{6.29}$$

O parâmetro α será determinado de tal modo que a expressão de $\bar{F}$ coincida aproximadamente com a expressão de F, dispensando-se assim o conhecimento das derivadas parciais f_1' e f_2' que constam da definição da função $F(t,y)$ em (6.27).
Para o feito, considere-se a combinação linear[2](ver Figura 6.6), de parâmetro $w \neq 0$,

$$\begin{aligned} F(t,y) &= f(t,y) + w\,(\bar{F}(t,y) - f(t,y)) \\ &= (1-w)\, f(t,y) + w\, \bar{F}(t,y)\ . \end{aligned} \tag{6.30}$$

Atendendo a (6.29), tem-se

$$\begin{aligned} F(t,y) &= (1-w)\, f(t,y) + w\, f(t,y) + \alpha\, h\, w\, f_1'(t,y) + \alpha\, h\, w\, f_2'(t,y) + \mathcal{O}(w\,(\alpha\, h)^2) \\ &= f(t,y) + \alpha\, h\, w\, f_1'(t,y) + \alpha\, h\, w\, f_2'(t,y) + \mathcal{O}(w\,(\alpha\, h)^2)\ . \end{aligned} \tag{6.31}$$

Comparando os termos contendo as derivadas parciais em (6.27) com os termos correspondentes de (6.31), concluimos que o parâmetro w deverá ser escolhido de

[2]Compare com o método SOR, pág 158.

modo que

$$\alpha\, h\, w = \frac{h}{2} \iff w = \frac{1}{2\,\alpha}, \quad \text{logo} \quad 1 - w = 1 - \frac{1}{2\,\alpha}\ .$$

Por conseguinte, de (6.30) e (6.31) resulta a aproximação,

$$\bar{F}(t,y) = \left(1 - \frac{1}{2\,\alpha}\right)\, f(t,y) + \frac{1}{2\,\alpha}\, \left(f(t + \alpha\, h, y + \alpha\, h\, f(t,y)\right).$$

Em conclusão, assumindo que $y(t_i) = y_i$ e , e após substituição em (6.26) de F por $\bar{F}$, obtém-se a seguinte família de métodos de segunda ordem, dependente do parâmetro $\alpha \neq 0$,

$$y_{i+1} = y_i + h\, \left[(1 - \frac{1}{2\,\alpha}) f(t_i, y_i) + \frac{1}{2\,\alpha}\, \left(f(t_i + \alpha\, h, y + \alpha\, h\, f(t_i, y_i))\right)\right]. \tag{6.32}$$

Nos próximos parágrafos analisaremos alguns casos particulares de métodos da família (6.32).

6.4.1 Método de Heun

Substituindo o parâmetro α em (6.32) por $\alpha = 1$, obtém-se o *método de Heun*[3]

$$y_{i+1} = y_i + \frac{h}{2}\, \left[f(t_i, y_i) + f(t_i + h, y_i + h\, f(t_i, y_i))\right]. \tag{6.33}$$

Interpretação geométrica

Na Figura 6.7 é dada uma interpretação geométrica deste método.

Uma vez que a função $f(t,y)$ define um *campo de direcções*, o ponto estimado y_{i+1} do método de Heun resulta de considerar a *média* dos declives $v_1 = f(t_i, y_i)$ e $v_2 = f(t_i + h, B)$, onde $B = y_i + h\, v_1$, das rectas tangentes à curva integral passando respectivamente nos pontos (t_i, y_i) e $(t_i + h, B)$.

6.4.2 Método do ponto médio ou Euler modificado

Substituindo o parâmetro α em (6.32) por $\alpha = 1/2$, obtém-se o *método do ponto médio* ou *Euler modificado*,

$$y_{i+1} = y_i + h\, f\left(t_i + \frac{h}{2}, y_i + \frac{h}{2}\, f(t_i, y_i)\right). \tag{6.34}$$

Na Fig. 6.8 é dada uma interpretação geométrica.

[3]Karl Heun, 1859-1929, matemático alemão.

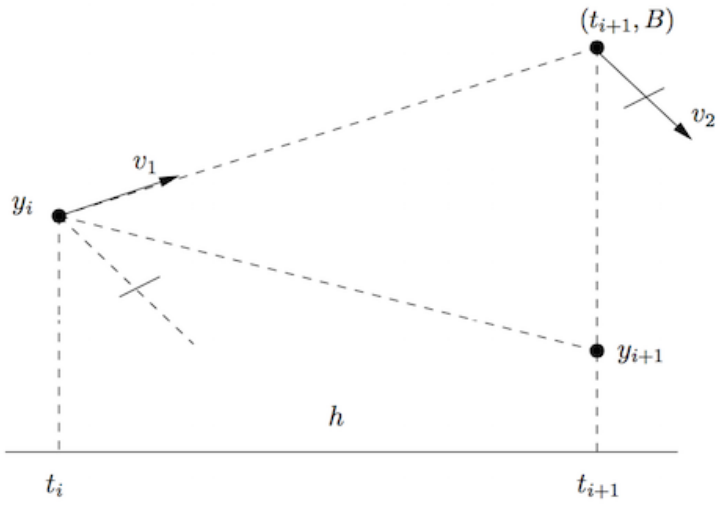


Figura 6.7: Método de Heun.

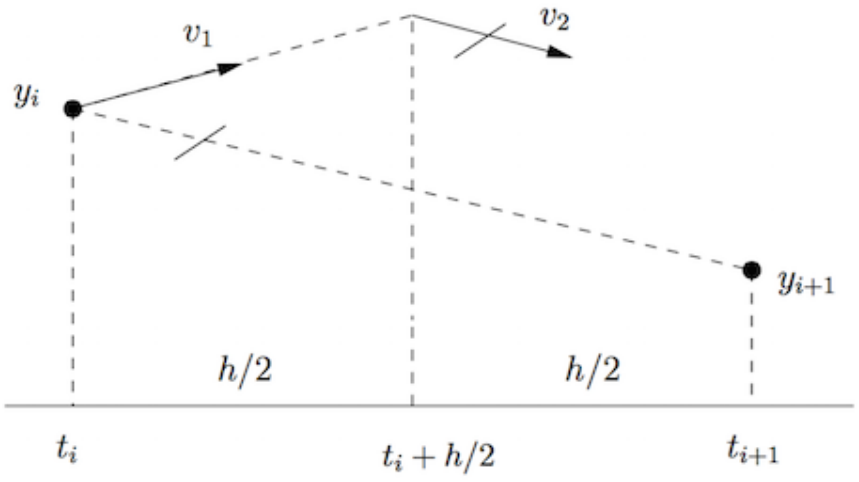


Figura 6.8: Método do ponto médio.

6.5 Método de Runge - Kutta de quarta ordem clássico

O método de Runge-Kutta[4] clássico é relativamente simples, oferecendo a vantagem relativamente aos métodos anteriores de possuir um erro de truncatura global da ordem de h^4. Ele pode ser deduzido generalizando o que se fez para os métodos de segunda ordem, embora a álgebra envolvida seja complicada. Obtém-se uma média pesada de 4 valores do campo de direcções $f(t, y)$, respectivamente à esquerda, ao centro e à direita no intervalo $[t_i, t_{i+1}]$. É costume denotar esses valores por v_1 a v_4:

$$
\begin{aligned}
&v_1 = f(t_i, y_i) \\
&v_2 = f(t_i + \frac{h}{2}, y_i + \frac{h}{2}\, v_1) \\
&v_3 = f(t_i + \frac{h}{2}, y_i + \frac{h}{2}\, v_2) \\
&v_4 = f(t_i + h, y_i + h\, v_3), \\
&\text{sendo a fórmula recursiva dada por,} \\
&y_{i+1} = y_i + h\, \frac{(v_1 + 2\, v_2 + 2\, v_3 + v_4)}{6}\ .
\end{aligned}
\tag{6.35}
$$

Na Figura 6.9 é dada interpretação geométrica para este método.

Note-se que no caso do campo de direcções não depender de y, isto é, quando $f(t, y) = f(t)$,

$$
\begin{aligned}
v_1 &= f(t_i) \\
v_2 = v_3 &= f(t_i + h/2) \\
v_4 &= f(t_i + h).
\end{aligned}
$$

Logo,

$$
y_{i+1} = y_i + \frac{h}{6}\left[f(t_i) + 4\, f(t_i + h/2) + f(t_i + h)\right]\ .
$$

Da expressão anterior concluimos que $y_{i+1} - y_i$ é uma aproximação do integral $\int_{t_i}^{t_{i+1}} f(t)\, dt$, mediante aplicação da regra de Simpson, pág. 250. Ora, sabemos que o erro de quadratura para esta regra, fixado o intervalo $[a, b] = [t_0, T]$ e o passo $h = (T - t_0)/N$, é da ordem $\mathcal{O}(h^4)$, confirmando-se assim indirectamente

[4]Carl David Runge, 1856-1927, matemático e físico, e M. Wilhelm Kutta, 1867-1944, matemático, ambos alemães.

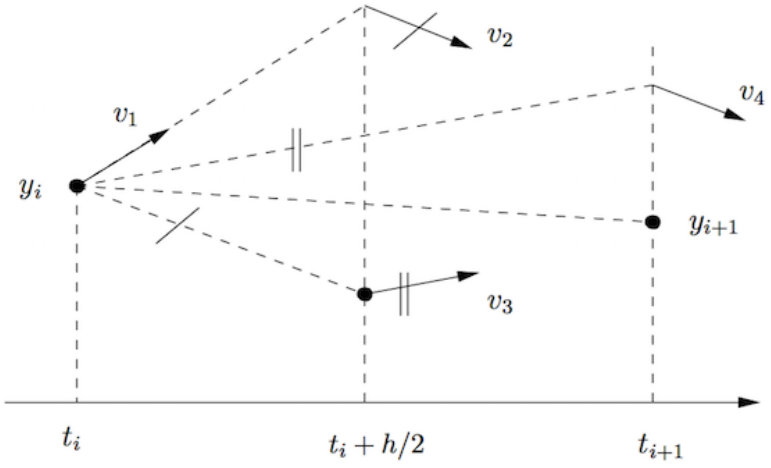


Figura 6.9: Método de Runge-Kutta clássico.

ser a ordem de convergência do método de Runge-Kutta clássico igualmente de quarta ordem.

No Exemplo 6.5 a seguir são comparados os métodos de segunda ordem de Heun, do ponto médio, e de Taylor, com o método de Runge-Kutta clássico de quarta ordem, num problema de valor inicial de solução conhecida, no intervalo $[0, 1]$. Utiliza-se respectivamente o passo $h = 0.2$ e o passo $h = 0.1$.

A partir das tabelas de valores calculados para cada um dos métodos referidos podemos confirmar numericamente a respectiva *ordem de convergência*, comparando o erro global em $x = 1$, para o passo 0.1, com o erro global nesse ponto, para o passo 0.2. Como se sabe, num método de segunda ordem o quociente desses erros deverá ser aproximadamente $1/4$, enquanto que num método de quarta ordem esse quociente deve ser próximo de $1/16$.

Exemplo 6.5. *Considere o problema de valor inicial*

$$\begin{cases} y'(x) = y(x) - x^2 + 1, & 0 \le x \le 1 \\ y(0) = 0.5, \end{cases}$$

cuja solução é

$$y(x) = 1 + 2\,x + x^2 - e^x/2\ .$$

Obtenha uma aproximação de $y(1)$, aplicando os métodos abaixo nomeados, com passo $h = 0.2$.
Usando um programa apropriado, repita os métodos referidos em (a),(b) e (c) a seguir, com passo $h = 0.1$.

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.2	0.826	0.829299	0.00329862
0.4	1.20692	1.21409	0.00716765
0.6	1.63724	1.64894	0.0116982
0.8	2.11024	2.12723	0.0169938
1.	2.61769	2.64086	0.0231715

Tabela 6.4: Método de Heun para o Exemplo 6.5, com $h = 0.2$.

Compare o respectivo erro em $x = 1$, e conclua sobre a ordem de convergência desses métodos.
(a) Método de Heun.

(b) Método do ponto médio.

(c) Método de Taylor de ordem dois.

(d) Método de Runge-Kutta de ordem quatro.

(a) Como o campo de direcções é definido por $f(x, y) = y - x^2 + 1$, tem-se:

$$\begin{array}{ll} v_1 = f(x,y) & = y - x^2 + 1 \\ B = y + h\, v_1 & = (1+h)\, y - h\, x^2 + h \\ v_2 = f(x+h, B) & = (1+h)\, y - (h+1)x^2 - 2\, h\, x + h - h^2 + 1 \\ v_1 + v_2 & = (2+h)\, y - (2+h)\, x_i^2 - 2\, h\, x + h - h^2 + 2 \ . \end{array}$$

O método é definido pela função $\Psi(x, y) = y + h/2\,(v_1 + v_2)$, donde a fórmula recursiva,

$$y_{i+1} = \left(1 + \frac{h}{2}\,(2+h)\right)\, y_i - \frac{h}{2}\,(2+h)\, x_i^2 - h^2\, x_i + \frac{h}{2}\,(h - h^2 + 2) \ . \qquad (6.36)$$

Para $h = 0.2$, o número de passos a realizar será $N = 1/h = 5$, e levando em consideração a aproximação inicial em $x = 0$, de (6.36) obtém-se,

$$\begin{array}{l} y_0 = 0.5 \\ y_{i+1} = 1.22\, y_i - 0.22\, x_i^2 - 0.04\, x_i + 0.216, \qquad i = 0 : (N-1) \ . \end{array} \qquad (6.37)$$

Na Tabela 6.4 encontra-se o resultado da aplicação de (6.37). Foram calculados os valores das aproximações sucessivas da solução y_i, bem como os respectivos erros $e_i = y(x_i) - y_i$. A Tabela 6.5 mostra resultados análogos quando reduzimos o passo a metade ou seja, para $N = 10$.

Dado que

$$\frac{|e_{10}|}{|e_5|} = \frac{0.0060618}{0.0231715} \simeq 0.262 \simeq 26\ \%,$$

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.1	0.657	0.657415	0.000414541
0.2	0.828435	0.829299	0.000863621
0.3	1.01372	1.01507	0.00134992
0.4	1.21221	1.21409	0.00187631
0.5	1.42319	1.42564	0.00244583
0.6	1.64588	1.64894	0.00306174
0.7	1.8794	1.88312	0.00372751
0.8	2.12278	2.12723	0.0044468
0.9	2.37497	2.3802	0.00522352
1.	2.6348	2.64086	0.0060618

Tabela 6.5: Método de Heun para o Exemplo 6.5, com $h = 0.1$.

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.2	0.828	0.829299	0.00129862
0.4	1.21136	1.21409	0.00272765
0.6	1.64466	1.64894	0.0042814
0.8	2.12128	2.12723	0.00594531
1.	2.63317	2.64086	0.00769233

Tabela 6.6: Método do ponto médio para o Exemplo 6.5, com $h = 0.2$.

confirmamos numericamente que o método é de segunda ordem de convergência, pois ao reduzirmos o passo h a metade, o erro global de truncatura é aproximadamente reduzido de 1/4.

(b) Usando agora a fórmula recursiva (6.34), é fácil concluir que para $h = 0.1$ se tem,

$$\begin{aligned} &y_0 = 0.5 \\ &y_{i+1} = 1.22\, y_i - 0.22 x_i^2 - 0.04\, x_i + 0.218 \qquad i = 0 : (N-1) \ . \end{aligned} \tag{6.38}$$

Na Tabela 6.6 mostram-se os resultados para este passo, e na Tabela 6.7 os valores calculados com passo $h = 0.1$. Dado que

$$\frac{|e_{10}|}{|e_5|} = \frac{0.00198065}{0.00769233} \simeq 0.257 \simeq 26\ \%,$$

concluimos de novo que o método é de segunda ordem de convergência para a solução do p.v.i. dado. No entanto, uma vez que, para o passo $h = 0.1$ o erro absoluto em $x = 1.0$ do método do ponto médio é inferior ao erro absoluto para o

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.1	0.65725	0.657415	0.000164541
0.2	0.828961	0.829299	0.000337371
0.3	1.01455	1.01507	0.000518415
0.4	1.21338	1.21409	0.000707491
0.5	1.42474	1.42564	0.000904288
0.6	1.64783	1.64894	0.00110834
0.7	1.8818	1.88312	0.001319
0.8	2.12569	2.12723	0.0015354
0.9	2.37844	2.3802	0.00175642
1.	2.63888	2.64086	0.00198065

Tabela 6.7: Método do ponto médio para o Exemplo 6.5, com $h = 0.1$.

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.2	0.83	0.829299	-0.000701379
0.4	1.2158	1.21409	-0.00171235
0.6	1.65208	1.64894	-0.0031354
0.8	2.13233	2.12723	-0.00510318
1.	2.64865	2.64086	-0.00778683

Tabela 6.8: Método de Taylor para o Exemplo 6.5, com $h = 0.2$.

método de Heun, concluimos que para o problema em causa o método do ponto médio produz melhores resultados numéricos do que o método de Heun.

Compare-se agora os resultados anteriores com o método de Taylor de segunda ordem.

(c) Como

$$\begin{aligned} f(x,y) &= y - x^2 + 1 \\ f_1'(x,y) &= -2\,x \\ f_2'(x,y) &= 1, \end{aligned}$$

e o método de Taylor de segunda ordem resulta da função

$$\Psi(x,y) = y + h\,f(x,y) + h^2/2\,\left((f_1'(x,y) + f_2'(x,y)\,f(x,y)\right),$$

obtém-se a seguinte fórmula recursiva para este método, com passo $h = 0.1$,

$$\begin{aligned} &y_0 = 0.5 \\ &y_{i+1} = 1.22\,y_i - 0.22\,x_i^2 - 0.04\,x_i + 0.22 \qquad i = 0 : (N-1)\ . \end{aligned} \tag{6.39}$$

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.1	0.6575	0.657415	-0.000085459
0.2	0.829487	0.829299	-0.000188879
0.3	1.01538	1.01507	-0.000313091
0.4	1.21455	1.21409	-0.000461324
0.5	1.42628	1.42564	-0.000637252
0.6	1.64979	1.64894	-0.000845062
0.7	1.88421	1.88312	-0.00108951
0.8	2.12861	2.12723	-0.001376
0.9	2.38191	2.3802	-0.00171067
1.	2.64296	2.64086	-0.00210049

Tabela 6.9: Método de Taylor para o Exemplo 6.5, com $h = 0.1$.

A partir das Tabelas 6.8 e 6.9, conclui-se que

$$\frac{|e_{10}|}{|e_5|} = \frac{0.00210049}{0.00778683} \simeq 0.270 \simeq 27\ \%,$$

o que sugere tratar-se de um método de segunda ordem, como seria de esperar. Comparando o erro global em $x = 1.0$, visto que $|e_{10}| \simeq 0.00210$ para o método de Taylor, $|e_{10}| \simeq 0.00606$ para o método de Heun, e $|e_{10}| \simeq 0.00198$ para o método do ponto médio, concluímos que embora estes erros sejam aproximadamente iguais, acontece neste exemplo que o método do ponto médio é ligeiramente mais preciso do que os dois restantes.

(d) Para aplicarmos o método de Runge-Kutta clássico, comecemos por determinar as expressões das 4 direcções do campo f que definem o método,

$$\begin{aligned}
v_1 &= f(x, y) = y - x^2 + 1 \\
v_2 &= f(x + h/2, y + h/2\, v1) = y + h/2\, v_1 - (x + h/2)^2 + 1 \\
v_3 &= f(x + h/2, y + h/2\, v_2) = y + h/2\, v_2 - (x + h/2)^2 + 1 \\
v_4 &= f(x + h, y + h\, v_3) = y + h\, v_3 - (x + h/2)^2 + 1 \ .
\end{aligned}$$

Substituindo nas expressões anteriores x e y, respectivamente por x_i e y_i, obtém-se, para $h = 0.1$,

$$\begin{aligned}
y_{i+1} &= y_i + \ (v_1(x_i, y_i) + 2\, v_2(x_i, y_i) + 2\, v_3(x_i, y_i) + v_4(x_i, y_i)) \\
&= 0.218593 - 0.0428\, x_i - 0.2214\, x_i^2 + 1.2214\, y_i \ .
\end{aligned}$$

Das Tabelas 6.10 e 6.11, conclui-se que

$$\frac{|e_{10}|}{|e_5|} = \frac{2.36159 \times 10^{-6}}{0.000036393} \simeq 0.065 \simeq 7\ \%,$$

Notando que $1/16 \simeq 0.0625$, o quociente de erros anterior confirma tratar-se de método de quarta ordem, como se esperava. ♦

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.2	0.829293	0.829299	5.28759×10^{-6}
0.4	1.21408	1.21409	0.0000114405
0.6	1.64892	1.64894	0.0000185828
0.8	2.1272	2.12723	0.0000268508
1.	2.64082	2.64086	0.000036393

Tabela 6.10: Método de Rung-Kutta clássico para o Exemplo 6.5, com $h = 0.2$.

x_i	y_i	$y(x_i)$	$e_i=y(x_i)-y_i$
0	0.5	0.5	0.
0.1	0.657414	0.657415	1.65962×10^{-7}
0.2	0.829298	0.829299	3.44923×10^{-7}
0.3	1.01507	1.01507	5.37779×10^{-7}
0.4	1.21409	1.21409	7.45476×10^{-7}
0.5	1.42564	1.42564	9.69002×10^{-7}
0.6	1.64894	1.64894	1.20939×10^{-6}
0.7	1.88312	1.88312	1.46771×10^{-6}
0.8	2.12723	2.12723	1.74508×10^{-6}
0.9	2.3802	2.3802	2.04264×10^{-6}
1.	2.64086	2.64086	2.36159×10^{-6}

Tabela 6.11: Método de Runge-Kutta clássico para o Exemplo 6.5, com $h = 0.1$.

6.6 Problemas de valor inicial para sistemas

Sistemas de $k \geq 2$ equações diferenciais ordinárias, de primeira ordem, são muito comuns nas aplicações. Tais sistemas podem ocorrer sob a forma

$$\begin{array}{l} y_1'(t) = f_1(t, y_1(t), y_2(t), \ldots, y_k(t)) \\ y_2'(t) = f_2(t, y_1(t), y_2(t), \ldots, y_k(t)) \\ \quad \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t_0 \leq t \leq T \\ y_k'(t) = f_k(t, y_1(t), y_2(t), \ldots, y_k(t)), \end{array} \tag{6.40}$$

dadas k condições iniciais $y_1(t_0) = \alpha_1$, $y_2(t_0) = \alpha_2$, $\cdots$, $y_k(t_0) = \alpha_k$.

Por exemplo, ao considerarmos uma certa equação diferencial de ordem k,

$$u^{(k)}(t) = \phi(t, u, u', \cdots, u^{k-1}), \qquad t_0 \leq t \leq T, \tag{6.41}$$

com k condições iniciais $u(t_0) = \alpha_1$, $u'(t_0) = \alpha_2$, $\cdots$, $u^{(k-1)}(t_0) = \alpha_k$, a equação (6.41) pode rescrever-se na forma de um sistema do tipo (6.40).
De facto, sejam

$$\begin{array}{l} y_1(t) = u(t) \\ y_2(t) = u'(t) \\ \quad \vdots \\ y_k(t) = u^{(k-1)}(t) \ . \end{array}$$

Derivando as igualdades anteriores, obtém-se o seguinte sistema de equações de primeira ordem,

$$\begin{array}{l} y_1'(t) = f_1(t, y_1(t), y_2(t), \ldots, y_k(t)) = y_2(t) \\ y_2'(t) = f_2(t, y_1(t), y_2(t), \ldots, y_k(t)) = y_3(t) \\ \quad \vdots \\ y_k'(t) = f_k(t, y_1(t), y_2(t), \ldots, y_k(t)) = u^{(k)}(t) = \phi(t, y_1(t), y_2(t), \ldots, y_k(t)), \end{array} \tag{6.42}$$

com k condições iniciais $y_1(t_0) = \alpha_1$, $y_2(t_0) = \alpha_2$, ..., $y_k(t_0) = \alpha_k$.

O sistema (6.42) traduz-se vectorialmente na forma

$$\begin{array}{l} y'(t) = F(t, y(t)) \\ y(t_0) = (\alpha_1, \alpha_2, \ldots, \alpha_k), \end{array} \tag{6.43}$$

a qual é formalmente idêntica e generaliza o problema de valor inicial (6.1), pág. 285. A função F caracteriza o *campo de velocidades* associado ao sistema de equações dado.

O Teorema 6.1, pág. 286, pode generalizar-se para sistemas do tipo (6.42).

Os métodos numéricos que estudámos podem facilmente ser adaptados para problemas de valor inicial como (6.43). Por exemplo, o método de Euler explícito aplicado ao sistema (6.43) dá origem à equação vectorial às diferenças,

$$y_{n+1} = y_n + h\, F(t, y_n), \qquad n = 0, 1, \ldots, N,$$

onde a função F tem por componentes as funções f_i definidas pelas expressões em (6.42)).

No exemplo a seguir aplicamos o método de Euler para resolver uma equação diferencial de segunda ordem, reduzindo-a a um sistema do tipo (6.42).

Exemplo 6.6. *Considere-se a equação diferencial de segunda ordem,*

$$u''(t) = 1 + t^2 + t\, u'(t), \qquad 0 \le t \le 1$$

e as condições iniciais

$$u(0) = 1, \qquad u'(0) = 2\ .$$

(a) Pretende-se aplicar o método de Euler para aproximar $u(1)$ e $u'(1)$, com passo desde $h = 0.2$ a $h = 0.025$ por bissecções sucessivas do passo 0.2. Sabe-se que a solução do problema dado toma os valores $u(1) = 4.08141$ e $u'(1) = 5.11881$ (com 6 algarismos significativos). Para cada um dos valores de h referidos, calcular as iteradas correspondentes do método de Euler,

$$\begin{array}{l} y_{1,k} \simeq u(t_k) \\ y_{2,k} \simeq u'(t_k), \qquad k = 0 : N \end{array}$$

dando uma tabela contendo os valores calculados para $t = 1$, bem como a norma $||y - y_{aprox}||_\infty$, sendo y o vector da solução exacta no ponto $t = 1$, e y_{aprox} o vector resultando da aplicação do método.
Qual é a ordem de convergência sugerida pelos valores obtidos?

(b)Traçar o gráfico das aproximações de $u(t)$ e $u'(t)$, para $0 \le t \le 1$, utilizando o passo $h = 0.01$.

(a) Fazendo $y_1(t) = u(t)$ e $y_2(t) = u'(t)$, o problema proposto pode escrever-se como um sistema de duas equações diferenciais de primeira ordem,

$$\begin{array}{l} y_1'(t) = y_2(t) = f_1(t, y_1, y_2) \\ y_2'(t) = u''(t) = 1 + t^2 + t\, y_2(t) = f_2(t, y_1, y_2), \qquad 0 \le t \le 1, \end{array}$$

sujeito às condições iniciais $y_{1,0} = 1$ e $y_{2,0} = 2$.

Fixado $h > 0$, as respectivas equações às diferenças para o método de Euler (6.5), pág. 289, escrevem-se

$$\left\{ \begin{array}{l} y_{1,n+1} = y_{1,n} + h\, f_1(t_n, y_{1,n}, y_{2,n}) = y_{1,n} + h y_{2,n} \\ y_{2,n+1} = y_{2,n} + h\, f_2(t_n, y_{1,n}, y_{2,n}) = y_{2,n} + h\,(1 + t_n^2 + t_n\, y_{2,n}), \qquad n = 0, 1, \ldots, \end{array} \right. \tag{6.44}$$

h	$y_{1,n}$	$y_{2,n}$	$\|\|y - y_n\|\|_\infty$
0.2	3.63494	4.44716	0.671653
0.1	3.83230	5.75474	0.364074
0.05	3.94945	4.92870	0.190113
0.025	4.01344	5.02159	0.0972245

Tabela 6.12: Aproximações em $t = 1$. O vector da solução exacta é $y(1) = (y_1(1), y_2(1))$ e $y_n = (y_{1,n}, y_{2,n})$ representa o vector calculado pelo método de Euler (ver Exemplo 6.6).

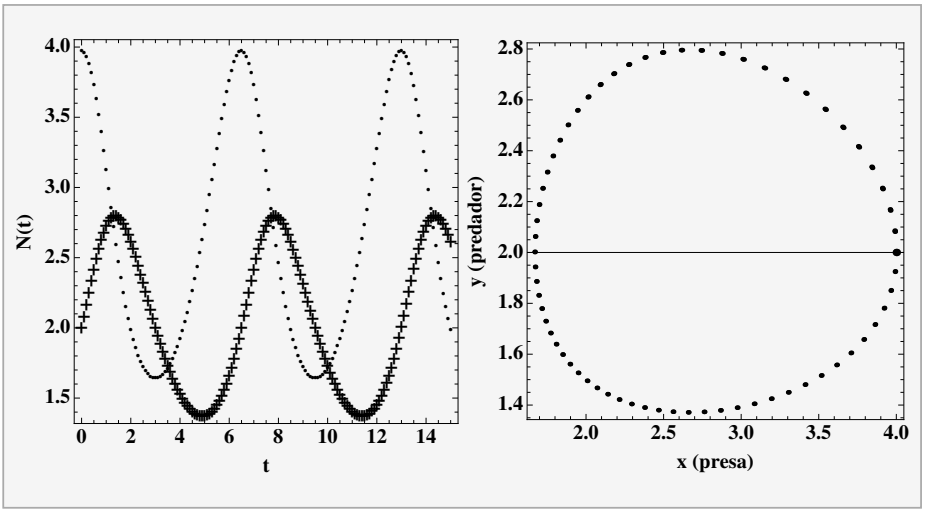


Figura 6.10: Método de Runge-Kutta clássico, com passo $h = 0.1$ (ver Exemplo 6.7).

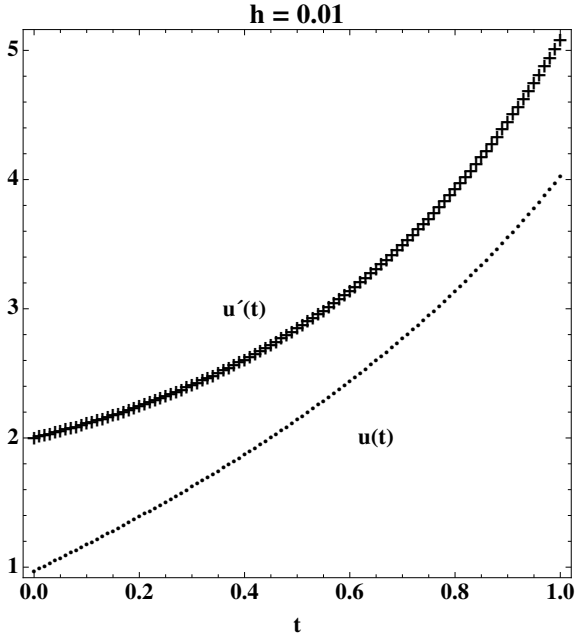


Figura 6.11: Aproximações de $u(t)$ e $u'(t)$ pelo método de Euler, com passo $h = 0.01$ (ver Exemplo 6.6).

com $y_{1,0} = 1$ e $y_{2,0} = 2$, e $t_n = n\,h$, para $n = 0 : 9$.

Na Tabela 6.12 apresentam-se os resultados obtidos para $t = 1.0$, respectivamente com passo $h = 0.2$, $h = 0.1$, $h = 0.05$ e $h = 0.025$.
Os quocientes dos valores calculados para a norma $||y - y_n||_\infty$,

$$\frac{0.364074}{0.671653} \simeq 0.542, \quad \frac{0.190113}{0.364074} \simeq 0.522, \quad \frac{0.0972245}{0.190113} \simeq 0.511,$$

sugerem que o quociente do erro global satisfaz

$$||e_{h/2}||_\infty \simeq \frac{1}{2}\,||e_h||_\infty \ .$$

A relação anterior indica que o método de Euler vectorial, neste exemplo, possui ordem de convergência um, de acordo com o que se viu no caso escalar.

(b) Na Figura 6.11 mostram-se os gráficos das aproximações calculadas para $u(t) = y_1(t)$, e $u'(t) = y_2(t)$, com $0 \le t \le 1$, para o passo $h = 0.01$. ♦

Exemplo 6.7. *Um modelo clássico permitindo o estudo da evolução de duas populações de animais, é conhecido pela designação de sistema de Lotka-Volterra*[5]*. Trata-se de um sistema de duas equações diferenciais de primeira ordem, do tipo*

$$\begin{aligned} x'(t) &= x(t)\,(r - \alpha\,y(t)) \\ y'(t) &= y(t)\,(-s + \beta\,x(t)), \end{aligned}$$

[5]Alfred J. Lotka, 1880 -1949, biofísico americano. Vito Volterra, 1860 - 1940, matemático italiano.

onde r, s, α e β são parâmetros positivos caracterizando as populações em causa. As incógnitas são as funções $x(t)$, representando o número de indivíduos habitualmente designados por "presas", e $y(t)$, representando o número de "predadores", num certo instante t (para um estudo detalhado destes sistemas ver, por exemplo, [33]).

Em particular, considere-se o sistema

$$\begin{aligned} x'(t) &= 1.2\,x(t) - 0.6\,x(t)\,y(t) \\ y'(t) &= -0.8\,y(t) + 0.3\,x(t)\,y(t), \qquad 0 \le t \le 15\ . \end{aligned} \tag{6.45}$$

Admita que o número inicial de presas é $x(0) = 4$, e que o número de predadores é $y(0) = 2$. Interessa-nos decidir se a população em causa apresenta ou não uma evolução periódica ao longo do espaço de tempo considerado.

Para o efeito, vamos adaptar o método de Runge-Kutta clássico (6.35), pág. 308, ao caso de sistemas com duas equações diferenciais (a generalização a sistemas com mais equações é igualmente simples).

Fixado o passo $h = 0.1$, na Fig. 6.10 são mostradas as aproximações calculadas por aplicação do método. No gráfico à esquerda $N(t)$ representa o número de indivíduos de cada uma das populações $x(t)$, e $y(t)$ e, no gráfico à direita encontra-se traçada uma curva constituída pelos pontos de coordenadas calculadas $(x(t), y(t))$, curva essa habitualmente designada por *retrato de fase* da solução do sistema diferencial.

Os cálculos foram refeitos com passo $h = 0.01$, e os resultados estão representados na Figura 6.12. A simulação numérica efectuada sugere que a população em causa evolui, de facto, de modo periódico ao longo do tempo.

Vejamos como se escrevem as equações às diferenças do método de Runge-Kutta aplicado ao sistema (6.45). A partir das equações diferenciais dadas, defina-se o campo de velocidades (f_1, f_2), onde

$$\begin{aligned} f_1(t, y_1, y_2) &= 1.2\,y_1 - 0.6\,y_1\,y_2 \\ f_2(t, y_1, y_2) &= -0.8\,y_2 + 0.3\,y_1\,y_2\ . \end{aligned}$$

Dado o passo $h > 0$, o método de Runge-Kutta (6.35), pág. 308, aplicado ao presente sistema de 2 equações, traduz-se nas seguintes expressões, para $i = 0 : (N-1)$, e $N = (T - t_0)/h$.

$$\begin{aligned} v_{1,1} &= f_1(t_i, y_{1,i}, y_{2,i}) = 1.2\,y_{1,i} - 0.6\,y_{1,i}\,y_{2,i} \\ v_{1,2} &= f_2(t_i, y_{1,i}, y_{2,i}) = -0.8\,y_{2,i} + 0.3\,y_{1,i}\,y_{2,i}. \end{aligned}$$

$$\begin{aligned} v_{2,1} &= f_1(t_i + h/2, y_{1,i} + h/2\,v_{1,1}, y_{2,i} + h/2\,v_{1,2}) \\ &= 1.2\,(y_{1,i} + h/2\,v_{1,1}) - 0.6\,(y_{1,i} + h/2\,v_{1,1})\,(y_{2,i} + h/2\,v_{1,2}) \end{aligned}$$

$$\begin{aligned} v_{2,2} &= f_2(t_i + h/2, y_{1,i} + h/2\,v_{1,1}, y_{2,i} + h/2\,v_{1,2}) \\ &= -0.8\,(y_{1,i} + h/2\,v_{1,1}) + 0.3\,(y_{1,i} + h/2\,v_{1,1})\,(y_{2,i} + h/2\,v_{1,2})\ . \end{aligned}$$

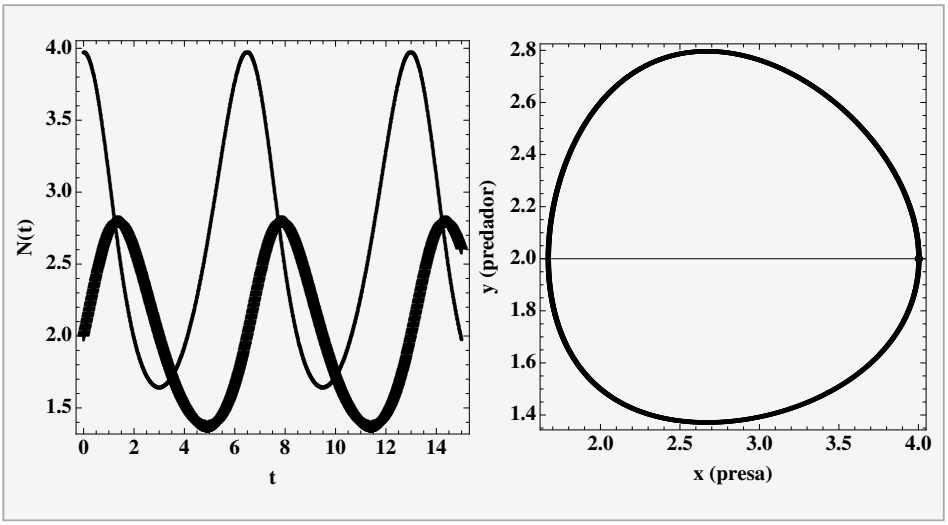


Figura 6.12: Método de Runge-Kutta clássico, com passo $h = 0.01$ (ver Exemplo 6.7).

$$\begin{aligned} v_{3,1} &= f_1(t_i + h/2, y_{1,i} + h/2\, v_{2,1}, y_{2,i} + h/2\, v_{2,2}) \\ &= 1.2\,(y_{1,i} + h/2\, v_{2,1}) - 0.6\,(y_{1,i} + h/2\, v_{2,1})\,(y_{2,i} + h/2\, v_{2,2}) \end{aligned}$$

$$\begin{aligned} v_{3,2} &= f_2(t_i + h/2, y_{1,i} + h/2\, v_{2,1}, y_{2,i} + h/2\, v_{2,2}) \\ &= -0.8\,(y_{2,i} + h/2\, v_{2,2}) + 0.3\,(y_{1,i} + h/2\, v_{2,1})\,(y_{2,i} + h/2\, v_{2,2})\ . \end{aligned}$$

$$\begin{aligned} v_{4,1} &= f_1(t_i + h, y_{1,i} + h\, v_{3,1}, y_{2,i} + h\, v_{3,2}) \\ &= 1.2\,(y_{1,i} + h\, v_{3,1}) - 0.6\,(y_{1,i} + h\, v_{3,1})\,(y_{2,i} + h\, v_{3,2}) \end{aligned}$$

$$\begin{aligned} v_{4,2} &= f_2(t_i + h, y_{1,i} + h\, v_{3,1}, y_{2,i} + h\, v_{3,2}) \\ &= -0.8\,(y_{2,i} + h\, v_{3,2}) + 0.3\,(y_{1,i} + h\, v_{3,1})\,(y_{2,i} + h\, v_{3,2})\ . \end{aligned}$$

Finalmente,

$$y_{1,i+1} = y_{1,i} + \frac{h}{6}\,[v_{1,1} + 2\, v_{2,1} + 2\, v_{3,1} + v_{4,1}]$$

$$y_{2,i+1} = y_{2,i} + \frac{h}{6}\,[v_{1,2} + 2\, v_{2,2} + 2\, v_{3,2} + v_{4,2}], \qquad i = 0 : (N-1)\ .$$

♦

6.7 Exercícios resolvidos

Exercício 6.1. *Considere o problema de valor inicial*

$$\begin{aligned} y'(t) &= \frac{t^2}{1-y(t)^2}, \qquad 0 \le t \le 1 \\ y(0) &= 0 \ . \end{aligned}$$

Embora não se conheça uma expressão para a solução do problema dado, sabe-se ([6], pág. 31) que a respectiva solução $y(t)$ satisfaz a equação

$$y^3(t) - 3\,y(t) + t^3 = 0, \tag{6.46}$$

como facilmente se pode verificar.

A solução $y(t)$ do p.v.i. está definida implicitamente através da equação (6.46)*. Esta circunstância de se conhecer a solução de um p.v.i. sob forma implícita é muito frequente nas aplicações. Os métodos numéricos que estudámos em capítulos anteriores são imprescindíveis na resolução de problemas desta natureza.*

Como ilustração, neste exercício iremos recorrer ao método de Newton para estimar (com pelo menos 10 dígitos significativos) a solução do p.v.i. proposto, num suporte discreto. Construiremos o polinómio interpolador desse suporte, o qual nos dará uma aproximação da solução no intervalo $[0,1]$*. Num determinado ponto deste intervalo, iremos comparar um valor obtido mediante aplicação de* 5 *iterações do método de Newton, com a estimativa da solução do problema, nesse ponto, calculada através do método de Heun* (6.33)*, pág. 306.*

(a) Fazendo $t = 1/5 = 0.2$, utilize a equação (6.46) *e o método de Newton, a fim de aproximar $y(0.2)$, com erro absoluto inferior a 10^{-10}.*

(b) Para $h = 0.2$, repita o processo da alínea anterior, para obter uma tabela $\{t_i, \bar{y}_i\}_{i=0}^{5}$, sendo $t_i = i\,h$ e $\bar{y}_i$ a última iterada calculada pelo método de Newton, a qual aproxima a solução do p.v.i. no ponto t_i. A partir dessa tabela construa o respectivo polinómio interpolador de grau 5*.*

(c) Para o passo $h = 0.2$, obtenha uma aproximação de $y(0.2)$ mediante aplicação do método de Heun. Determine o respectivo erro aproximado, usando um valor da tabela referida na alínea (b).

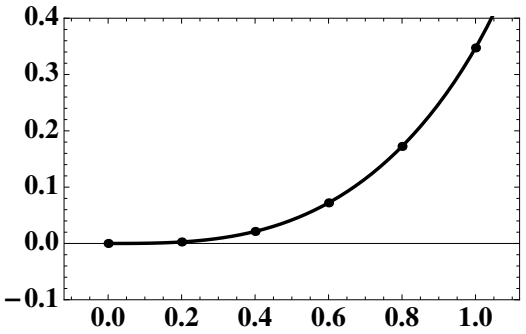
(a) Vejamos que é possível aplicar o método de Newton, garantindo convergência quadrática, e erro inferior à tolerância dada.

Fixado $t \in (0,1]$, seja F o seguinte polinómio, na variável y,

$$F(y) = y^3 - 3\,y + t, \qquad y \in I = [0, 1/2] \ .$$

A equação $F(y) = 0$ tem uma única raiz z no intervalo I. Com efeito,

$$\begin{aligned} &F(0) \times F(1/2) < 0 \\ &F'(y) = 3\,(y^2 - 1) < 0 \quad \forall\, y \in I, \end{aligned}$$



0	0
$\frac{1}{5}$	0.00266667298770
$\frac{2}{5}$	0.0213365711528
$\frac{3}{5}$	0.0721250654662
$\frac{4}{5}$	0.172373901567
1	0.347296355334

Figura 6.13: Aproximações de $y(t_i)$ para 5 iterações do método de Newton.

pelo que existe um só zero de F em I. Além disso,

$$F''(y) = 6\,y \geq 0, \qquad \forall\, y \in I.$$

Assim, como $F(0) = t > 0$ e $F''(y) \geq 0$, tomando para aproximação inicial de z, o valor $y^{(0)} = 0$, podemos garantir convergência (quadrática) monótona do método de Newton.

Vamos estimar quantas iterações serão necessárias para garantir a precisão requerida para a última iterada que efectuarmos.

Fazendo

$$\mathcal{K} = \frac{1}{2}\,\frac{\max\,|F''(y)|}{\min_{y\in I}\,|F'(y)|} = \frac{1}{2}\,\frac{F''(1/2)}{|F'(0)|} = \frac{1}{2},$$

sabemos que, para cada iterada $y^{(i)}$, é satisfeita a desigualdade

$$\begin{aligned} |z - y^{(i)}| \;&\leq \frac{1}{\mathcal{K}}\;\left(\mathcal{K}\,|z - y^{(0)}|\right)^{2^i}, \quad i = 1, 2, \ldots \\ &< \frac{1}{\mathcal{K}}\;\left(\frac{\mathcal{K}}{2}\right)^{2^i} = 2\,\left(\frac{1}{4}\right)^{2^i}, \end{aligned}$$

onde a última desigualdade é válida, uma vez que $|z - y^{(0)}| < 1/2$. Por conseguinte, fazendo $\epsilon = 10^{-10}$, se impusermos a condição,

$$\begin{aligned} &2\left(\frac{1}{4}\right)^{2^i} < \epsilon \iff 2^i\,\ln(1/4) < \ln(\epsilon/2) \\ &\Leftrightarrow i > \frac{\ln(\epsilon/2)/\ln(1/4)}{\ln(2)} \simeq 4.1, \end{aligned}$$

concluimos que para $i \geq 5$, o erro absoluto da iterada $y^{(5)}$ é inferior a ϵ.

Para $y^{(0)} = 0$, e $t = 1/5$, aplicando o método de Newton $y^{(i+1)} = y^{(i)} - F(y^{(i)})/F'(y^{(i)})$, para $i = 0:5$, obtêm-se os seguintes valores:

i	$y^{(i)}$
0	0
1	0.00266666666667
2	0.00266667298770
3	0.00266667298770
4	0.00266667298770
5	0.00266667298770

Assim, o valor $\bar{y} = 0.00266667298770$, aproxima o valor da solução $y(1/5)$, com erro inferior a 10^{-10}.

(b) A Figura 6.13 mostra uma tabela contendo o resultado de 5 iterações do método de Newton, partindo de $y^{(0)} = 0$, respectivamente para $t = 1/5$, até $t = 1$, por acréscimos de $h = 1/5$. Na mesma figura encontra-se traçado o polinómio interpolador $p_5(t)$ dessa tabela,

$$\begin{aligned} p_5(t) \quad &= 0.0052964408\,t - 0.056569983\,t^2 + 0.539683984\,t^3 - \\ &-0.314165012\,t^4 + 0.173050926\,t^5. \end{aligned}$$

Uma verificação da "proximidade" do polinómio interpolador $p_5(t)$, relativamente à solução $y(t)$ do p.v.i. dado, pode fazer-se substituindo a expressão do polinómio na equação diferencial, e considerar a função

$$E(t) = p_5'(t) - t^2/(1 - p_5'^2(t)), \qquad 0 \leq t \leq 1\ .$$

No gráfico da Figura 6.14 está representada a função E anterior. Este gráfico mostra-nos que o polinómio interpolador que determinámos, aproxima a solução do p.v.i. com um erro absoluto global inferior a 0.02. Por conseguinte, se o problema concreto subjacente ao modelo matemático que constitui o p.v.i. dado, for tal que um erro global de 0.02 possa ser considerado aceitável, então o nosso polinómio interpolador pode ser considerado como "solução" do problema proposto no intervalo $[0, 1]$. Caso contrário, poderiámos refinar a "malha" e usar de novo o método de Newton para um suporte de espaçamento menor. A partir da tabela de valores calculados construiríamos o respectivo polinómio interpolador, aumentando o grau de interpolação. Propomos ao leitor que efectue essas experiências numéricas.

(c) Seja

$$f(t,y) = \frac{t^2}{1-y^2} \implies f[t+h, y+h\,f(t,y)] = \frac{(t+h)^2}{1 - \left(y + \dfrac{h\,t^2}{1-y^2}\right)^2}\ .$$

No método de Heun é utilizada a equação às diferenças,

$$y_{i+1} = y_i + \frac{h}{2}\,[f[t_i, y_i) + f(t_i + h, y_i + h\,f(t_i, y_i))], \quad i = 0, 1, \ldots\ .$$

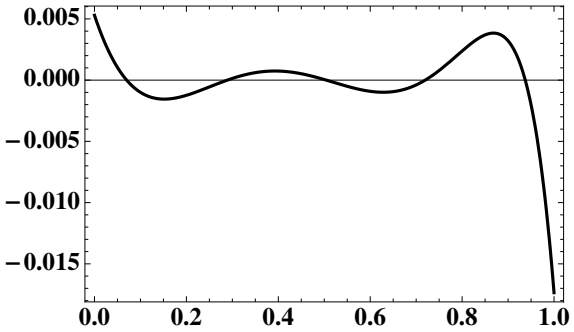


Figura 6.14: Função $E(t) = p_5'(t) - t^2/(1 - p_5'^2(t))$.

Aplicando ao p.v.i. em causa, resulta

$$y_0 = 0$$
$$y_{i+1} = y_i + \frac{h}{2}\left[\frac{t_i^2}{1-y_i^2} + \frac{(t_i+h)^2}{1-\left(y + \dfrac{h\,t_i^2}{1-y_i^2}\right)^2}\right], \quad i = 0, 1, \ldots \ .$$

Assim, para $t_0 = 0$ e $h = 1/5 = 0.2$, obtém-se,

$$y(0.2) \simeq y_1 = \frac{0.2}{2}\,0.2^2 = 0.004 \ .$$

Comparando com o valor $y^{(5)}$, da tabela de iteradas do método de Newton para $t = 1/5$, conclui-se que o erro de y_1 calculado pelo método de Heun é

$$y(0.2) - y_1 \simeq 0.00266667298770 - 0.004 = -0.00133333 \ .$$

♦

O Exercício 6.2 a seguir ilustra o ganho de precisão que é em geral possível obter quando se substitui um método de primeira ordem de convergência por outro de maior ordem. Por exemplo, no problema proposto, apesar de usarmos o método de Runge-Kutta com um passo h valendo o dobro do passo utilizado para o método de Euler, o resultado calculado para o primeiro método é cerca de três vezes mais preciso do que o resultado para o segundo.

Exercício 6.2. *Considere o problema de valor inicial*

$$\begin{aligned} y_1'(t) &= y_1(t) - 4\,y_2(t) \\ y_2'(t) &= -y_1(t) + y_2(t) \\ y_1(0) = 1, &\quad y_2(0) = 0 \ . \end{aligned}$$

cuja solução é $y(t) = (y_1(t), y_2(t))$*, onde*

$$y_1(t) = \frac{e^{-t} + e^{3\,t}}{2}, \qquad y_2(t) = \frac{e^{-t} - e^{3\,t}}{4} \ .$$

Pretende-se obter valores aproximados da solução, em $t = 0.2$.

(a) Aplicando o método de Euler explícito, com passo $h = 0.1$.

(b) Idem, utilizando o método de Runge-Kutta de quarta ordem para sistemas, com passo $h = 0.2$.

(c) Em cada caso, comparar o número de algarismos significativos da aproximação $\bar{y} = (\bar{y}_1, \bar{y}_2)$*, usando a função*

$$Sig(\bar{y}) = |\log_{10}(||y - \bar{y}||_\infty)|,$$

onde $y = (y_1(0.2), y_2(0.2))$.

(a) O campo de velocidades associado ao p.v.i. dado é da forma

$$F(t, y_1, y_2) = (f_1(t, y_1, y_2), f_2(t, y_1, y_2)) = (y_1 - 4\,y_2, -y_1 + y_2).$$

As equações às diferenças para o método de Euler, escrevem-se

$$\begin{array}{l} y_{1,n+1} = y_{1,n} + h\, f_1(t_n, y_{1,n}, y_{2,n}) = y_{1,n} + h\,(y_{1,n} - 4\,y_{2,n}) \\ y_{2,n+1} = y_{2,n} + h\, f_2(t_n, y_{1,n}, y_{2,n}) = y_{2,n} + h\,(-y_{1,n} + y_{2,n}), \qquad n = 0, 1, \ldots \ . \end{array}$$

Como o ponto onde se pretende aproximar a solução é $T = 0.2$, o número de passos a efectuar é $N = (T - 0)/h = 2$, onde $h = 0.1$, e o vector inicial é $y_0 = (y_{1,0}, y_{2,0}) = (1, 0)$.

Primeiro passo $\quad t_0 = 0$

$$\begin{array}{l} y_{1,1} = y_{1,0} + h\,(y_{1,0} - 4\,y_{2,0}) = 1 + 0.1 \times 1 = 1.1 \\ y_{2,1} = y_{2,0} + h\,(-y_{1,0} + y_{2,0}) = 0 + 0.1 \times (-1) = -0.1 \end{array}$$

Segundo passo $\quad t_1 = 0 + h = 0.1$

$$\begin{array}{l} y_{1,2} = y_{1,1} + h\,(y_{1,1} - 4\,y_{2,1}) = 1.1 + 0.1 \times (1.1 + 0.4) = 1.25 \\ y_{2,2} = y_{2,1} + h\,(-y_{1,1} + y_{2,1}) = -0.1 + 0.1\,(-1.1 - 0.1) = -0.22 \ . \end{array}$$

Assim, a aproximação pretendida em $t = 0.2$, é

$$\bar{y} = (\bar{y}_{1,2}, \bar{y}_{2,2}) = (1.25, -0.22) \ .$$

Como a solução do p.v.i. nesse ponto vale (com 12 algarismos significativos),

$$y = (y_{1,2}, y_{2,2}) = (1.32042477673, -0.250847011828),$$

tem-se,

$$||y - \bar{y}||_\infty = 0.0704248 \quad \text{e} \quad Sig(\bar{y}) = |\log(0.0704248)| \simeq 1.15 \ .$$

Ou seja, a aproximação $\bar{y}$ possui aproximadamente um algarismos significativo (para a norma $||.||_\infty$).

(b) Neste caso, para $h = 0.2$, a aproximação da solução obtida através do método de Runge-Kutta será calculada num só passo.

As fórmulas de passagem do ponto $t = t_n$ ao ponto $t = t_{n+1}$ são as seguintes:

$$\begin{cases} v_{1,1} = f_1(t_n, y_{1,n}, y_{2,n}) = y_{1,n} - 4\, y_{2,n} \\ v_{1,2} = f_2(t_n, y_{1,n}, y_{2,n}) = -y_{1,n} + y_{2,n} \ . \end{cases}$$

$$\begin{cases} v_{2,1} = f_1(t_n + h/2, y_{1,n} + h/2\, v_{1,1}, y_{2,n} + h/2\, v_{1,2}) = y_{1,n} + \dfrac{h}{2}\, v_{1,1} - 4\,(y_{2,n} + \dfrac{h}{2}\, v_{1,2}) \\ v_{2,2} = f_2(t_n + h/2, y_{1,n} + h/2\, v_{1,1}, y_{2,n} + h/2\, v_{1,2}) = -(y_{1,n} + \dfrac{h}{2}\, v_{1,1}) + y_{2,n} + \dfrac{h}{2}\, v_{1,2} \end{cases}$$

$$\begin{cases} v_{3,1} = f_1(t_n + h/2, y_{1,n} + h/2\, v_{2,1}, y_{2,n} + h/2\, v_{2,2}) = y_{1,n} + \dfrac{h}{2}\, v_{2,1} - 4\,(y_{2,n} + \dfrac{h}{2}\, v_{2,2}) \\ v_{3,2} = f_2(t_n + h/2, y_{1,n} + h/2\, v_{2,1}, y_{2,n} + h/2\, v_{2,2}) = -(y_{1,n} + \dfrac{h}{2}\, v_{2,1}) + y_{2,n} + \dfrac{h}{2}\, v_{2,2} \end{cases}$$

$$\begin{cases} v_{4,1} = f_1(t_n + h, y_{1,n} + h\, v_{3,1}, y_{2,n} + h\, v_{3,2}) = y_{1,n} + h\, v_{3,1} - 4\,(y_{2,n} + h\, v_{3,2}) \\ v_{4,2} = f_2(t_n + h, y_{1,n} + h\, v_{3,1}, y_{2,n} + h\, v_{3,2}) = -(y_{1,n} + h\, v_{3,1}) + y_{2,n} + h\, v_{3,2} \ . \end{cases}$$

Finalmente,

$$y_{1,n+1} = y_{1,n} + \frac{h}{6}(v_{1,1} + 2\, v2, 1 + 2\, v_{3,1} + v_{4,1})$$

$$y_{2,n+1} = y_{2,n} + \frac{h}{6}(v_{1,2} + 2\, v2, 2 + 2\, v_{3,2} + v4, 2), \qquad n = 0, 1 \ldots \ .$$

Para $t_0 = 0$, e aproximações iniciais $y_{1,0} = 1$ e $y_{2,0} = 0$, obtém-se

$$\begin{cases} v_{1,1} = y_{1,0} - 4\, y_{2,0} = 1 \\ v_{1,2} = -y_{1,0} + y_{2,0} = -1 \ . \end{cases}$$

$$\begin{cases} v_{2,1} = (y_{1,0} + 0.1\, v_{1,1}) - 4\,(y_{2,0} + 0.1\, v_{1,2}) = 1.1 - 4\,(-0.1) = 1.5 \\ v_{2,2} = -1.1 + (-0.1) = -1.2 \end{cases}$$

$$\begin{cases} v_{3,1} = (y_{1,0} + 0.1\, v_{2,1}) - 4\,(y_{2,0} + 0.1\, v_{2,2}) = 1.15 - 4\,(-0.12) = 1.63 \\ v_{3,2} = -1.15 - 0.12 = -1.27 \end{cases}$$

$$\begin{cases} v_{4,1} = (y_{1,0} + 0.2\, v_{3,1}) - 4\,(y_{2,0} + 0.2\, v_{3,2}) = 1.326 - 4\,(-0.254) = 2.342 \\ v_{4,2} = -1.326 - 0.254 = -1.580 \ . \end{cases}$$

Donde,

$$y_{1,1} = 1 + \frac{0.2}{6}(1 + 2 \times 1.5 + 2 \times 1.63 + 2.342) = 1.320066667$$

$$y_{2,1} = 0 + \frac{0.2}{6}(-1 + 2 \times (-1.2) + 2 \times (-1.27) - 1.580) = -0.250666666 \ .$$

Assim, a aproximação pretendida em $t = 0.2$, é

$$\bar{y} = (\bar{y}_{1,1}, \bar{y}_{2,1}) = (1.320066667, -0.250666666),$$

e

$$||y - \bar{y}||_\infty = ||(0.00035811, -0.000180345)||_\infty = 0.00035811 \ .$$

Logo,

$$Sig(\bar{y}) = |\log(0.00035811)| \simeq 3.45 \ .$$

Ou seja, a aproximação $\bar{y}$ possui mais do que três algarismos significativos (para a norma $||.||_\infty$).

♦

6.8 Leituras aconselhadas

W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, John Wiley & Sons, New York, 1992, Ch. 8.

M. Braun, *Differential Equations and Their Applications*, Springer, New York, 1993.

S. C. Chapra, R. P. Canale, *Métodos Numéricos para Engenharia*, McGraw–Hill, São Paulo, 2008.

A. McAndrew, Exploring Runge-Kutta formulas with a computer algebra system, *Electronic J. of Mathematics and Technology*, 10, 84-100 (2016).

D. A. Sanchez, R. C. Allen Jr. and W. T. Kyner, *Differential Equations*, Addison-Wesley, Massachusetts, 1988, Ch. 7.

Apêndice A

Suplemento: testes e exames resolvidos

São aqui apresentados alguns enunciados de testes e exames (acompanhados da sua resolução), propostos nos últimos anos aos alunos frequentando a cadeira de Matemática Computacional, disciplina comum a diversos cursos do Instituto Superior Técnico.

A seguir estão reunidas as fórmulas essenciais introduzidas ao longo do curso.

A.1 Formulário

Teoria de erros e representação de números

Erro absoluto e erro relativo

$$
\begin{array}{l}
x, \tilde{x} \in \mathbb{R}, \quad x \approx \tilde{x} \\
e_{\tilde{x}} = x - \tilde{x}, \qquad \delta_{\tilde{x}} = \dfrac{e_{\tilde{x}}}{x}, \qquad x \neq 0
\end{array}
$$

$$
\begin{array}{ll}
\text{erro absoluto :} & |e_{\tilde{x}}| \\
\text{erro relativo :} & |\delta_{\tilde{x}}|, \qquad x \neq 0
\end{array}
$$

Erros de arredondamento

$$x = \sigma(0.a_1a_2...)_\beta \beta^t,\ a_1 \neq 0; \quad \tilde{x} = fl(x) \in FP(\beta, n, t_1, t_2)$$

$$
\begin{array}{l}
|e_{\tilde{x}}| \leq \beta^{t-n}, \quad |\delta_{\tilde{x}}| \leq \beta^{1-n} := \mu_c \quad \text{(arredondamento por corte)} \\
|e_{\tilde{x}}| \leq \dfrac{1}{2}\beta^{t-n}, \quad |\delta_{\tilde{x}}| \leq \dfrac{1}{2}\beta^{1-n} := \mu_s \quad \text{(arredondamento simétrico)}
\end{array}
$$

Propagação de erros

$$x, \tilde{x} \in \mathbb{R}^n, \quad x \approx \tilde{x}$$

$$e_{f(\tilde{x})} = f(x) - f(\tilde{x}) \approx \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) e_{\tilde{x}_k}$$

$$\delta_{f(\tilde{x})} = \frac{e_{f(\tilde{x})}}{f(x)} \approx \sum_{k=1}^{n} p_{f,k}(x)\delta_{\tilde{x}_k}, \qquad p_{f,k}(x) = \frac{x_k \frac{\partial f}{\partial x_k}(x)}{f(x)}$$

$$\delta_{\tilde{f}(\tilde{x})} \approx \sum_{k=1}^{n} p_{f,k}(x)\delta_{\tilde{x}_k} + \sum_{k=1}^{m} q_k(x)\delta_{\mathrm{arr}_k}$$

Métodos iterativos para equações não lineares

Estimativas de erro

$$e_k = z - x_k \simeq -\frac{f(x_k)}{f'(x_k)}$$

$$|e_k| \leq \frac{|f(x_k)|}{\min_{a \leq x \leq b} |f'(x)|}$$

Método da bissecção

$$x_{k+1} = \frac{a_k + b_k}{2}, \quad f(a_k)f(b_k) < 0$$

$$|x - x_{k+1}| \leq |x_{k+1} - x_k|, \qquad |x - x_k| \leq \frac{b-a}{2^k}$$

Método de Newton

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

$$x - x_{k+1} = -\frac{f''(\xi_k)}{2f'(x_k)}(x - x_k)^2, \qquad \xi_k \in int(x_k, z)$$

$$|x - x_k| \leq \frac{1}{\mathbb{K}} \left(\mathbb{K}|x - x_0|\right)^{2^k}$$

$$e_k = x - x_k \simeq x_{k+1} - x_k$$

Método da secante

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$$

$$x - x_{k+1} = -\frac{f''(\xi_k)}{2f'(\eta_k)}(x - x_k)(x - x_{k-1}),$$
$$\eta_k \in int(x_{k-1}, x_k), \qquad \xi_k \in int(x_{k-1}, x_k, z)$$

$$|x - x_{k+1}| \leq \mathbb{K}\,|x - x_k|\,|x - x_{k-1}|, \quad \mathbb{K} = \frac{\max|f''|}{2\min|f'|}$$

Método do ponto fixo

$x_{k+1} = g(x_k)$

$$|x - x_{k+1}| \leq \frac{L}{1-L}|x_{k+1} - x_k|$$

$$|x - x_k| \leq L^k|x - x_0|, \qquad |x - x_k| \leq \frac{L^k}{1-L}|x_1 - x_0|$$

Normas e Condicionamento

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

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

$$\|\mathbf{A}\|_2 = (\rho(\mathbf{A}^T\mathbf{A}))^{1/2}$$

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$$

$$\|\delta_{\tilde{\mathbf{x}}}\| \leq \frac{\text{cond}(\mathbf{A})}{1 - \text{cond}(\mathbf{A})\,||\delta_{\tilde{\mathbf{A}}}\|} \,(\|\delta_{\tilde{\mathbf{A}}}\| + \|\delta_{\tilde{\mathbf{b}}}\|), \text{ sistema } \mathbf{Ax} = \mathbf{b}$$

Métodos iterativos para sistemas lineares

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{Cx} + \mathbf{d} \quad \rightarrow \quad \mathbf{x}^{(k+1)} = \mathbf{Cx}^{(k)} + \mathbf{d}$$

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|\mathbf{C}\|^k \|\mathbf{x} - \mathbf{x}^{(0)}\|, \qquad \|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{C}\|^k}{1 - \|\mathbf{C}\|}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \frac{\|\mathbf{C}\|}{1 - \|\mathbf{C}\|}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|$$

Método de Jacobi

$C = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}), \qquad x_i^{(k+1)} = (b_i - \sum_{j=1, j\neq i}^{n} a_{ij}\, x_j^{(k)})/a_{ii}$

Método de Gauss-Seidel

$$C = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$$
$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)})/a_{ii}$$

Método SOR

$$\mathbf{C} = -(\mathbf{L} + \omega^{-1}\mathbf{D})^{-1}(\mathbf{U} + (1 - \omega^{-1})\mathbf{D})$$
$$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} + \omega\mathbf{D}^{-1}(\mathbf{b} - \mathbf{Lx}^{(k+1)} - \mathbf{Ux}^{(k)})$$

Método de Newton para sistemas não lineares

$$\mathbf{J}(\mathbf{x}^{(k)})\Delta\mathbf{x}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)}) \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}$$

Interpolação polinomial

Interpolação de Lagrange

$$l_i(x) = \prod_{j=0, j\neq i}^{n} (\frac{x - x_j}{x_i - x_j}), \qquad p_n(x) = \sum_{i=0}^{n} y_i \, l_i(x)$$

Interpolação de Newton

$$\begin{cases} f[x_j] = f(x_j), \quad j = 0, ..., n \\ f[x_j, ..., x_{j+k}] = \dfrac{f[x_{j+1}, ..., x_{j+k}] - f[x_j, ..., x_{j+k-1}]}{x_{j+k} - x_j}, \quad j = 0, ..., n-k, \quad k = 1, ..., n \end{cases}$$

$$p_n(x) = f[x_0] + \sum_{i=1}^{n} f[x_0, ..., x_i](x - x_0) \cdots (x - x_{i-1})$$

$$e_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

Mínimos quadrados

$$\begin{bmatrix} \langle \phi_0, \phi_0 \rangle & \langle \phi_0, \phi_1 \rangle & \dots & \langle \phi_0, \phi_m \rangle \\ \langle \phi_1, \phi_0 \rangle & \langle \phi_1, \phi_1 \rangle & \dots & \langle \phi_1, \phi_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \phi_m, \phi_0 \rangle & \langle \phi_m, \phi_1 \rangle & \dots & \langle \phi_m, \phi_m \rangle \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \langle \phi_0, f \rangle \\ \langle \phi_1, f \rangle \\ \vdots \\ \langle \phi_m, f \rangle \end{bmatrix}$$

$$\langle \phi_i, \phi_j \rangle = \sum_{k=0}^{n} \phi_i(x_k)\phi_j(x_k), \qquad \langle \phi_i, f \rangle = \sum_{k=0}^{n} \phi_i(x_k) f_k$$

Integração numérica

Regra dos trapézios

$$T_N(f) = \frac{h}{2} \left[f(x_0) + f(x_N) + 2 \sum_{i=1}^{N-1} f(x_i) \right]$$

$$E_N^T(f) = -\frac{(b-a)\, h^2}{12} f''(\xi) \qquad \xi \in (a, b)$$

Regra de Simpson

$$S_N(f) = \frac{h}{3} \left[f(x_0) + f(x_N) + 4 \sum_{i=1}^{N/2} f(x_{2i-1}) + 2 \sum_{i=1}^{N/2-1} f(x_{2i}) \right]$$

$$E_N^S(f) = -\frac{(b-a)\, h^4}{180} f^{(4)}(\xi) \qquad \xi \in (a, b)$$

Métodos numéricos para equações diferenciais

Euler explícito

$$y_{i+1} = y_i + h\, f(t_i, y_i)$$

$$|y(t_i) - y_i| \leq \tfrac{h\,M}{2\,K}\left(e^{K\,(t_i - t_0)} - 1\right),$$
$$K = max_{t\in[t_0,t_i],y\in\mathbb{R}}\,|\frac{\partial f(t,y)}{\partial y}|, \quad M = max_{t\in[t_0,t_i]}\,|y''(t)|$$

Euler implícito

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1})$$

Taylor de ordem k

$$y_{i+1} = y_i + hf(t_i, y_i) + ... + \frac{h^k}{k!} f^{(k-1)}(t_i, y_i)$$

Métodos de Runge-Kutta de ordem 2

$$y_{i+1} = y_i + \left(1 - \frac{1}{2\alpha}\right) hf(t_i, y_i) + \frac{1}{2\alpha} hf(t_i + \alpha h, y_i + \alpha h f(t_i, y_i))$$

Método do ponto médio ($\alpha = 1/2$)

$$y_{i+1} = y_i + h\, f\left(t_i + h/2, y_i + h/2\, f(t_i, y_i)\right)$$

Método de Heun ($\alpha = 1$)

$$y_{i+1} = y_i + \frac{h}{2}\, f\left[f(t_i, y_i) + f(t_i + h, y_i + h\, f(t_i, y_i)\right]$$

Método de Runge-Kutta de ordem 4 clássico

$$V_1 = f(t_i, y_i)$$

$$V_2 = f(t_i + \frac{h}{2}, y_i + \frac{h}{2}\, V_1)$$

$$V_3 = f(t_i + \frac{h}{2}, y_i + \frac{h}{2}\, V_2)$$

$$V_4 = f(t_i + h, y_i + hV_3)$$

$$y_{i+1} = y_i + \frac{h}{6}(V_1 + 2V_2 + 2V_3 + V_4)$$

A.2 Testes e exames

A.2.1

[1.0] **1)** Sejam $x = 2$ e $y = e^{0.692}$. Sem calcular $fl(x) - fl(y)$, obtenha um majorante do erro relativo de arredondamento simétrico dessa diferença (expresso em percentagem). Admita que os cálculos são efectuados num sistema decimal de representação numérica com 4 dígitos na mantissa. Justifique.

2) Considere a sucessão $\{x_m\}$, tal que $x_0 = 2$ e

$$x_{m+1} = x_m - \frac{x_m^2 - a}{2\,x_m}, \qquad m = 0, 1, \ldots, \qquad a > 0 \ .$$

[1.0] **(a)** Admitindo que a sucessão converge, determine o seu limite. Justifique.

[1.0] **(b)** É ou não verdade que a sucessão em causa possui convergência linear? Justifique.

3) Para obter um valor aproximado da raiz z da equação $x^3 = x + 1$, situada no intervalo $I = [1, 2]$, pretende-se usar o método do ponto fixo, com uma das seguintes funções iteradoras:

$$g(x) = (x+1)^{1/3}, \qquad h(x) = \frac{2\,x^3 + 1}{3\,x^2 - 1}, \qquad r(x) = x^3 - 1 \ .$$

[1.0] **(a)** Diga, justificando, se alguma delas coincide com a função iteradora do método de Newton.

[1.0] **(b)** Se usar a função h e o ponto inicial $x_0 = 1.5$, poderá garantir convergência monótona? Justifique.

[1.0] **(c)** Sendo $x_0 \neq z$ e $x_0 \in I$, uma das funções iteradoras não poderá ser utilizada para aproximar o valor de z. Diga, justificando, que função é essa.

4) Considere o sistema linear $A\,x = b$, tal que

$$A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & a & 0 \\ 1 & -1 & 1 \end{bmatrix}, \quad a \in \mathbb{R}, \quad b = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} \ .$$

[0.5] **(a)** Diga, justificando, se existe algum valor do parâmetro a para o qual não seja possível factorizar a matriz A segundo o algoritmo de Doolittle ou de Crout.

[1.5] **(b)** Sendo $a = 0$, obtenha a factorização de Crout da matriz A. A partir dessa factorização descreva como resolveria o sistema $A\,x = b$ (não é necessário efectuar cálculos).

[1.5] **(c)** Para $a = 0$, diga se é ou não verdade que $cond(A)_\infty > 6$. Justifique.

[0.5] **(d)** A partir da definição de norma matricial induzida por norma vectorial, calcule $|| - \frac{1}{2}\,I||_1$, sendo I a matriz identidade (3×3).

(27 de Abril de 2010, MEEC)

Resolução

1. Sabe-se que

$$\delta_{fl(x)-fl(y)} \simeq \frac{x}{x-y}\,\delta_{fl(x)} - \frac{y}{x-y}\,\delta_{fl(y)} \ .$$

Atendendo a que 2 tem representação exacta no sistema, $fl(2) = 2$, tem-se que $\delta_{fl(x)} = 0$. Se μ_s designar a unidade de arredondamento simétrico, sabe-se que $\delta_{fl(y)} \leq \mu_s = 0.5 \times 10^{-3}$. Assim,

$$|\delta_{fl(x)-fl(y)}| \leq \frac{|y|}{|x-y|}\,u_s \simeq \frac{1.998}{0.002293} \times 0.5 \times 10^{-3} \simeq 0.436 \ .$$

2 a) A sucessão é gerada pela função iteradora, contínua em $\mathbb{R}^+$,

$$g(x) = x - \frac{x^2 - a}{2\,x} = \frac{x^2 + a}{2\,x} \ .$$

Seja α o limite dessa sucessão. Então,

$$\alpha = \lim_{m\to\infty} x_{m+1} = \lim_{m\to\infty} g(x_m) = g(\alpha) \ .$$

Logo, α é ponto fixo de g. Ora $g(x) = x \Leftrightarrow x^2 + a = 2\,x^2 \Leftrightarrow x^2 = a$, pelo que os pontos fixos de g são $\pm\sqrt{a}$. Como se parte de $x^{(0)} = 2 > 0$, o ponto fixo em causa é $\alpha = \sqrt{a}$.

2 b) Atendendo a que $g \in C^1(\mathbb{R}^+)$ e

$$g'(x) = \frac{1}{2}\frac{x^2 - a}{x^2} = \frac{1}{2}(1 - \frac{a}{x^2}),$$

resulta $g'(\sqrt{a}) = 0$, pelo que a convergência é supralinear (pelo menos quadrática).

3 a) Seja $f(x) = x^3 - x - 1 = 0$, e $z \in [1, 2]$ um zero de f. A função iteradora de Newton é

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^3 - x - 1}{3\,x^2 - 1} = \frac{2\,x^3 + 1}{3\,x^2 - 1} \ .$$

Assim, h é a função iteradora de Newton.

3 b) Atendendo a que $f \in C^2(A)$, $f(1.5) = 0.875 > 0$ e $f''(x) = 6\,x \quad \forall x \in A$, resulta que $f(1.5) \times f''(x) > 0$, $\forall x \in A$. Além disso, f' não muda de sinal em A. Sabe-se que estas condições são suficientes para garantir convergência monótona do método de Newton para o zero (único) $z \in A$ de f.

3 c) Como $x^3 = x + 1 \Leftrightarrow x = x^3 - 1 = r(x)$, o zero $z \in A$ é ponto fixo de r. No entanto, $r'(x) = 3\,x^2$, logo $|r'(z)| > 1$, pelo que z é repulsor para a função iteradora r. Assim, escolhido $x^{(0)}$ nas condições do enunciado, a sucessão $x_{k+1} = r(x_k)$, para $k = 0, 1, \ldots$, não pode convergir para z.

4 a) Dado que $det(A_1) = 1 \neq 0$, e $det(A_2) = -a + 1$, se $a + 1 = 0$, isto é, para $a = -1$ não é possível efectuar factorização de Doolittle ou de Crout.

4 b)

$$\begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} . \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$1^{\underline{a}}$ coluna de L e $1^{\underline{a}}$ linha de U:

$$\begin{array}{l} l_{11} = 1, \quad l_{21} = -1, \quad l_{31} = 1 \\ u_{12} = 1, \quad u_{13} = 1. \end{array}$$

$2^{\underline{a}}$ coluna de L e $2^{\underline{a}}$ linha de U:

$$\begin{array}{l} -1 + l_{22} = 0 \Rightarrow l_{22} = 1 \\ l_{31}\, u_{12} + l_{32} = -1 \Rightarrow l_{32} = -1 - 1 = -2 \\ l_{21}\, u_{13} + l_{22}\, u_{23} = 0 \Rightarrow u_{23} = -l_{21}\, u_{13}/l_{22} = 1. \end{array}$$

Entrada l_{33}:

$$1 - 2 + l_{33} = 1 \Leftrightarrow l_{33} = 2.$$

Assim,

$$A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = L.U \quad .$$

Para determinar a solução de $A\,x = b$, resolve-se primeiro o sistema triangular inferior $L\,g = b$ (por substituições para diante), e depois o sistema triangular superior $U\,x = g$ (por substituições para trás).

4 c) Sendo

$$A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix},$$

calcule-se A^{-1}:

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & -2 & 0 & -1 & 0 & 1 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 2 & 1 & 2 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1/2 & 1 & 1/2 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1/2 & -1 & -1/2 \\ 0 & 1 & 0 & 1/2 & 0 & -1/2 \\ 0 & 0 & 1 & 1/2 & 1 & 1/2 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1/2 & 0 & -1/2 \\ 0 & 0 & 1 & 1/2 & 1 & 1/2 \end{array}\right] = [I\,|A^{-1}] \ .$$

Por conseguinte,

$$||A||_\infty = max(3, 1, 3) = 3, \qquad ||A^{-1}||_\infty = max(1, 1, 2) = 2 \ .$$

Assim,

$$cond(A)_\infty = 3 \times 2 = 6,$$

pelo que a desigualdade é falsa.

4 d)

$$|| - \frac{1}{2}\, I||_1 = \frac{1}{2}\, ||I||_1 = \frac{1}{2},$$

visto que

$$||I||_1 = max_{||x||_1=1}\ ||I\, x||_1 = max_{||x||_1=1}\, ||x||_1 = 1\ .$$

A.2.2

Grupo I

Considere a equação

$$e^x - x^2 - 2\, x = 1/3$$

1) Mostre que a equação tem uma única raiz z_1 no intervalo $[0.5, 0.6]$. **[2.5]**

2) Para $n = 0, 1, \ldots$, considere as sucessões

$$(S1) \qquad x_{n+1} = \ln(x_n^2 + 2\, x_n + 1/3) = g_1(x_n)$$

$$(S2) \qquad w_{n+1} = \frac{e^{w_n} - w_n^2 - 1/3}{2} = g_2(w_n)\ .$$

a) Mostre que qualquer raiz positiva da equação é ponto fixo da função iteradora g_1 e reciprocamente. **[2.5]**

b) Sabendo que $\alpha = 2.36$ é uma aproximação de um ponto fixo da função g_1, pode garantir convergência local da sucessão $(S1)$ para este ponto fixo? **[2.5]**

3) Mostre que é possível obter aproximações da raiz z_1 usando a sucessão $(S2)$. Indique um intervalo onde poderá escolher a iterada inicial. **[2.5]**

4) Efectue duas iterações usando a sucessão $(S2)$, com $w_0 = 0.55$. Dê um majorante para o erro absoluto da aproximação obtida. **[2.5]**

5) Diga o que entende por ordem de convergência. Determine a ordem de convergência da sucessão $(S2)$, bem como uma aproximação do respectivo coeficiente assimptotico de convergência. **[2.5]**

Grupo II

6) Considere as matrizes **[2.5]**

$$B = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix} \quad \text{e} \quad x = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Determine $\|B\|_2$ e $\|B\, x\|_1$.

7) Obtenha a factorização de Doolittle da matriz **[2.5]**

$$A = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -4 & 5 & 0 \\ 0 & 0 & 1 & 5 \end{bmatrix}$$

e, a partir dela, calcule $Det(A)$.

(Exame 26 de Abril de 2007, LEC/LEGM)

Resolução

1) Sejam $f(x) = e^x - x^2 - 2\,x - 1/3 = 0$ e $I = [0.5, 0.6]$. Atendendo a que f é contínua, $f(0.5) \simeq 6.5 \times 10^{-2} > 0$ e $f(0.6) \simeq -7.1 \times 10^{-2} < 0$, sabemos (teorema de Bolzano) que existe pelo menos uma raiz da equação em I.
Ora, $f'(x) = e^x - 2\,x - 2$, $f''(x) = e^x - 2 = 0$ se e só se $x = \ln(2) \simeq 0.69 > 0.6$. Então, f' mantém sinal em I, pelo que z_1 é raiz única nesse intervalo.

2 a) Seja $z > 0$ tal que $e^z - z^2 - 2\,z = 1/3 \Leftrightarrow e^z = z^2 + 2z + 1/3 \Leftrightarrow z = \ln(z^2 + 2\,z + 1/3) = g_1(z)$. Ou seja, z é ponto fixo de g_1. Reciprocamente, se z é ponto fixo de g_1, então z é raiz da equação dada.

2 b) Para $\alpha = 2.36$, $g_1(\alpha) \simeq \alpha$. Como $g \in C^1(\mathbb{R}^+)$, $g_1'(x) = \dfrac{2\,x + 2}{x^2 + 2\,x + 1/3}$ e

$$0 < g_1'(\alpha) = \frac{2\,\alpha + 2}{\alpha^2 + 2\,\alpha + 1/3} \simeq 0.633 < 1\ .$$

Conclui-se que α é ponto fixo atractor da função g_1, pelo que se escolhermos um ponto inicial x_0 suficientemente próximo de α, a sucessão $(S1)$ converge para α.

3) Sendo $I = [0.5, 0.6]$ e $g_2(x) = \dfrac{e^x - x^2 - 1/3}{2} \in C^1(I)$, resulta $g_2'(x) = \dfrac{e^x - 2\,x}{2}$, $g_2''(x) = \dfrac{e^x - 2}{2}$. Como $g''(x) = 0$ se e só se $x = \ln(2) = 0.693 > 0.6$, resulta que g_2' é estritamente monótona em I. Ora, $g_2'(0.5) \simeq 0.32436$, $g'(0.6) \simeq 0.31106$, pelo que g_2 é função estritamente crescente. Assim,

$$0.5326\ldots = g_2(0.5) \leq g_2(x) \leq g_2(0.6) = 0.5643,$$

donde $g_2(I) \subset I$. Além disso,

$$L = max_{x \in I}\, |g_2'(x)\,| = g_2'(0.5) = 0.32436 < 1\ .$$

Atendendo ao teorema do ponto fixo, a raiz $z_1 \in I$ pode ser aproximada usando a sucessão $(S2)$.

4) Sabemos (alínea anterior) que $L = 0.32436$ e que

$$|z_1 - w_2| \leq \frac{L}{1 - L}\,|w_2 - w_1|\ .$$

A partir de $w_0 = 0.55$, obtém-se:

$$w_1 = g_2(w_0) = 0.5487098423 w_2 = g_2(w_1) = 0.5483012328 \ .$$

Assim,
$|z_1 - w_2| \leq 0.48008 \times 0.000409 \simeq 1.96 \times 10^{-4}$.

5) Dada uma sucessão de números reais $(x_n) \underset{n}{\longrightarrow} x$, se existir o limite dado a seguir (onde $p \geq 1$ e $k_\infty > 0$), dizemos que a sucessão possui ordem de convergência p (sendo k_∞ designado por coeficiente assimptótico de convergência):

$$\lim_{n\to\infty} \frac{|x - x_{n+1}|}{|x - x_n|^p} = k_\infty \ .$$

A sucessão $(S2)$ é gerada por $g_2 \in C^1(\mathbb{R})$. Sabe-se que $z_1 \simeq 0.5483$ (ver alínea anterior), donde

$$\lim_{n\to\infty} \frac{|z_1 - x_{n+1}|}{|z_1 - x_n|} = |g_2'(z_1)| \simeq g_2'(0.5483) = 0.317 \ .$$

Assim, a sucessão $(S2)$ possui ordem 1 e $k_\infty \simeq 0.317$.

6) Como B simétrica, $B^T B = B^2$ e $\rho(B^2) = (\rho(B))^2$.
Assim, $\|B\|_2 = [\rho(B^T B)]^{1/2} = \sqrt{\rho^2(B)} = \rho(B)$. Ora, $Det(\lambda I - B) = \begin{vmatrix} \lambda & 2 \\ 2 & \lambda \end{vmatrix} =$ 0 se e só se $\lambda = \pm 2$, donde $\|B\|_2 = \rho(B) = 2$. Como $B\,x = [-2 \quad 2]^T$, $\|B\,x\|_1 = 2 + 2 = 4$.

7) A matriz A é tridiagonal.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ 0 & l_{32} & 1 & 0 \\ 0 & 0 & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & 0 & 0 \\ 0 & u_{22} & u_{23} & 0 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -4 & 5 & 0 \\ 0 & 0 & 1 & 5 \end{bmatrix}$$

Cálculo das entradas de U e de L:

$$\begin{aligned} & u_{11} = 4, \quad u_{12} = -1 \\ & 4\,l_{21} = -1 \Leftrightarrow l_{21} = -1/4, \\ & 1/4 + u_{22} = 3 \Leftrightarrow u_{22} = 11/4 \\ & u_{23} = -1 \\ & 11/4\,l_{32} = -4 \Leftrightarrow l_{32} = -16/11, \\ & 16/11 + u_{33} = 5 \Leftrightarrow u_{33} = 39/11 \\ & u_{34} = 0, \\ & l_{43} \times 39/11 = 1 \Leftrightarrow l_{43} = 11/39 \\ & u_{44} = 5 \ . \end{aligned}$$

Por conseguinte,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/4 & 1 & 0 & 0 \\ 0 & -16/11 & 1 & 0 \\ 0 & 0 & 11/39 & 1 \end{bmatrix} \begin{bmatrix} 4 & -1 & 0 & 0 \\ 0 & 11/4 & -1 & 0 \\ 0 & 0 & 39/11 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

e

$$Det(A) = Det(L) \times Det(U) = 4 \times 11/4 \times 39/11 \times 5 = 195 \ .$$

A.2.3

1) Sabe-se que 1.9999 e 3.14 resultam de arredondamentos simétricos.

[2.0] **(a)** Estime o erro absoluto do valor de $\sin(1.9999 \times 3.14)$. Apresente todos os cálculos que efectuar.

[1.5] **(b)** Quantos algarismos significativos pode garantir para o valor mencionado na alínea anterior? Justifique.

[2.0] **(c)** Diga se a função $\Psi(a,b) = \sin(a \times b)$ é bem condicionada para pontos $(a,b) \neq (0,0)$, tais que $a \times b \simeq 2\,k\,\pi$, dado $k > 0$. Justifique a sua resposta começando por calcular o número de condição $P_{\Psi,1}(a,b)$.

2) Considere a equação $\cos(x) \times \cosh(x) = 1$ [onde $\cosh(x) = (e^x + e^{-x})/2$], a qual possui uma raiz (única) no intervalo $[4,5]$.

[1.5] **(a)** Diga, justificando, se poderá aplicar o método da bissecção para calcular uma aproximação de z, começando no intervalo $[4.5, 5]$.

[2.0] **(b)** Calcule o valor da iterada x_4 do método da bissecção, partindo de $x_0 = 4.7$ e $x_1 = 4.9$. Obtenha um majorante do erro relativo de x_4. Justifique.

[2.5] **(c)** Escolha um intervalo, um valor inicial x_0, e diga se pode garantir que a sucessão $(x_k)_{k\geq 0}$ obtida pelo método de Newton converge para o número z. No caso afirmativo poderá dizer que a convergência dessa sucessão é linear? Justifique.

[2.5] **(d)** Fazendo $x_0 = 4.75$, obtenha a segunda iterada do método de Newton, e estime o respectivo erro absoluto. Justifique.

3) Considere o sistema linear $A\,x = b$, sendo

$$A = \begin{bmatrix} -6 & 3 \\ 1 & -5 \end{bmatrix}, \qquad b = \begin{bmatrix} -3 \\ -4 \end{bmatrix} \ .$$

[2.0] **(a)** O número de condição da matriz A (para a norma $||\,.\,||_1$), é menor que $5/2$? Justifique.

[2.0] **(b)** Diga, justificando, se o método de Jacobi é convergente para a solução do sistema dado, caso inicie o processo usando $x^{(0)} = (100, -100)^T$.

[2.0] **(c)** Fazendo $x^{(0)} = (0,0)^T$, e efectuando cálculos exactos, obtenha a iterada $x^{(2)}$ bem como um majorante do respectivo erro (para a norma $||.||_\infty$) .

(Teste 15 de Abril 2011, MEEC)

Resolução

1(a) Seja $z = \Psi(a,b) = \sin(a \times b)$. Para $\bar{a} = 1.9999$ e $\bar{b} = 3.14$, aproximações obtidas por arredondamento simétrico respectivamente de a e b, sabe-se que os erros

absolutos satisfazem as desigualdades $|e_{\bar{a}}| \leq 0.5 \times 10^{-4}$ e $|e_{\bar{b}}| \leq 0.5 \times 10^{-2}$. Como $\bar{z} = \sin(\bar{a} \times \bar{b}) = -0.0034993$, utilizando a fórmula de propagação de erro da função Ψ,

$$e_{\bar{z}} \simeq \frac{\partial \Psi}{\partial a}(\bar{a}, \bar{b})\, e_{\bar{a}} + \frac{\partial \Psi}{\partial b}(\bar{a}, \bar{b})\, e_{\bar{b}},$$

obtém-se

$$e_{\bar{z}} \simeq \bar{b}\, \cos(\bar{a}\, \bar{b})\, e_{\bar{a}} + \bar{a}\, \cos(\bar{a}\, \bar{b})\, e_{\bar{b}}\ .$$

Atendendo a que $\cos(\bar{a}\, \bar{b}) \simeq 0.999$, resulta

$$\begin{aligned} |e_{\bar{z}}| &\leq 3.14 \times 0.999 \times 0.5 \times 10^{-4} + 1.9999 \times 0.999 \times 0.5 \times 10^{-2} \\ &\leq 0.00016 + 0.0099 \simeq 0.010 = 0.1 \times 10^{-1}\ . \end{aligned}$$

1(b) Visto que $\bar{z} = -0.34993 \times 10^{-2}$ e $|e_{\bar{z}}| \leq 0.1 \times 10^{-1}$ (ver alínea anterior), temos

$$|e_{\bar{z}}| \leq 0.1 \times 10^{-2-(-1)},$$

donde se pode concluir que $\bar{z}$ não possui nenhum algarismo significativo. De facto,

$$|\delta_{\bar{z}}| \simeq \frac{0.01}{0.0035} \simeq 2.9\ .$$

Ou seja, o erro relativo da aproximação é, aproximadamente, 290 %.

1(c) Atendendo a que

$$|P_{\psi,1}(a,b)| = |\frac{a\, \partial \Psi(a,b)/\partial a}{\Psi(a,b)}| = |\frac{a\, b\, \cos(a \times b)}{\sin(a \times b)}|,$$

para valores (a, b), com $a, b \neq 0$, tais que $\sin(a \times b) \simeq 0$, o numerador do membro à direita da expressão anterior é finito, mas o denominador é próximo de zero. Nessas condições $|P_{\psi,1}(a,b)| >> 1$ e a função é mal condicionada. Tal acontece, em particular, para valores de $a, b \neq 0$ tais que $a \times b \simeq 2\, k\pi$, com k inteiro.
Notar que o grande erro relativo do resultado obtido na alínea (b) deve-se ao facto da função ser mal condicionada numa região contendo pontos (a, b) próximos dos valores aproximados $(\bar{a}, \bar{b})$ utilizados nessa alínea.

2(a) Como f é contínua e $f(4.5) \times f(5) \simeq -210 < 0$, podemos aplicar o método da bissecção iniciando-o com o intervalo $J = [4.5, 5]$ considerado.

2(b) Seja $f(x) = \cos(x)\, (e^x + e^{-x}) - 2 = 0$.

$$x_2 = \frac{4.7 + 4.9}{2} = 4.8, \qquad f(4.7) < 0, \quad f(x_2) > 0, \quad \Rightarrow z \in [4.7, 4.8]$$

$$x_3 = \frac{4.7 + 4.8}{2} = 4.75, \quad f(x_3) > 0, \quad \Rightarrow z \in [4.7, 4.75]$$

$$x_4 = \frac{4.7 + 4.75}{2} = 4.725\ .$$

Então,

$$|z - x_4| \leq |x_4 - x_3| = 0.025\ .$$

Atendendo a que $z > 4.7$, resulta

$$\delta_{x_4} = \frac{|z - x_4|}{|z|} < \frac{0.025}{4.7} \simeq 0.053 \ .$$

2(c) Seja $f(x) = \dfrac{\cos x}{2}\,(e^x + e^{-x}) - 1$. Esta função é continuamente diferenciável, quantas vezes quanto se queira, em $\mathbb{R}$. Verifica-se que $f(4) \simeq -18.8$ e $f(5) \simeq 20.1$. Como f é contínua no intervalo $I = [4, 5]$ e muda de sinal nesse intervalo, pelo teorema de Bolzano conclui-se que a equação $f(x) = 0$ possui pelo menos uma raíz z em $(4, 5)$. Atendendo a que,

$$\begin{aligned} f'(x) &= \frac{-sin\, x\,(e^x + e^{-x}) + \cos x\,(e^x - e^{-x})}{2} \\ f''(x) &= -\sin x\,(e^x - e^{-x}), \end{aligned}$$

levando em conta que no intervalo I a função sin é negativa, e $(e^x - e^{-x}) > 0$, resulta que nesse intervalo f'' é positiva. Por conseguinte, f' é função estritamente crescente em I. Mas, $f'(4) \simeq 2.8 > 0$, donde se conclui que $f'(x) > 0\ \forall x \in I$. Assim, f é estritamente crescente no intervalo, pelo que o zero z é único. Por exemplo, em $I = [4.7, 4.9]$, sabemos que existe um único zero z de f. Se escolher, por exemplo, $x_0 = 4.9$, como $f(x_0) \times f''(x) > 0\ \forall x \in I$, sabe-se que o método converge para z, visto que $f \in C^2(I)$, muda de sinal nos extremos do intervalo, é aí estritamente crescente, com f'' positiva. Atendendo a que z é zero simples (pois $f'(z) \neq 0$), a convergência será quadrática.

2(d) Para $x_0 = 4.75$, obtém-se:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 4.73042215565$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 4.73004088772$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 4.73004074486,$$

donde

$$\begin{aligned} e_0 &\simeq x_1 - x_0 \simeq -0.02 \\ e_1 &\simeq x_2 - x_1 \simeq -0.00038 \\ e_2 &\simeq x_3 - x_2 \simeq -0.143 \times 10^{-6} \ . \end{aligned}$$

3(a)

$$||A||_1 = max(7, 8) = 8$$

$$A^{-1} = \frac{1}{det(A)} \begin{bmatrix} -5 & -3 \\ -1 & -6 \end{bmatrix} = \frac{1}{27} \begin{bmatrix} -5 & -3 \\ -1 & -6 \end{bmatrix}$$

$$||A^{-1}||_1 = max(6/27, 9/27) = 9/27 = 1/3 \ .$$

Logo, $cond_1(A) = ||A||_1\, ||A^{-1}|_1| = 8/3 > 5/2$, pelo que a desigualdade dada é falsa.

3(b) Dado que a matriz A é estritamente dominante por linhas, o método é convergente para $x = A^{-1}\,b$, qualquer que seja a aproximação inicial $x^{(0)}$.

3(c) Fórmulas computacionais do método:

$$\begin{aligned} x_1^{(k+1)} &= \frac{-3 - 3\,x_2^{(k)}}{-6} = \frac{1}{2} + \frac{1}{2}\,x_2^{(k)} \\ x_2^{(k+1)} &= \frac{-4 - x_1^{(k)}}{-5} = \frac{4}{5} + \frac{1}{5}\,x_1^{(k)} \qquad k = 0, 1, \ldots \ . \end{aligned}$$

Assim,

$$C_J = \begin{bmatrix} 0 & 1/2 \\ 1/5 & 0 \end{bmatrix} \qquad ||C_J||_\infty = max(1/2, 1/5) = 1/2 \ .$$

Como

$$x^{(1)} = (1/2, 4/5)^T \qquad x^{(2)} = (1/2 + 4/10, 4/5 + 1/10)^T = (9/10, 9/10)^T,$$

resulta $x^{(2)} - x^{(1)} = (4/10, 1/10)^T$ e $||x^{(2)} - x^{(1)}||_\infty = max(2/5, 1/10) = 2/5$. Por conseguinte,

$$||x - x^{(2)}||_\infty \leq \frac{||C_J||_\infty}{1 - ||C_J||_\infty}\,||x^{(2)} - x^{(1)}||_\infty \leq ||x^{(2)} - x^{(1)}||_\infty = 2/5 \ .$$

A.2.4

1) Considere o sistema de equações não lineares

$$\begin{cases} 4\,x_1 + x_2^3 + x_3 &= 7 \\ x_1\,x_3 + 5\,x_2 &= 1 \\ x_1^2 - x_2^2 + x_3^3 &= -5 \end{cases}$$

o qual possui uma solução $z = (z_1, z_2, z_3)$, em $D = [0, 3] \times [0, 3] \times [-2, 0]$. Pretende-se aproximar z aplicando o método de Newton, partindo de $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$.

(a) Diga se existe algum número real a, tal que o vector $x^{(0)} = (0, 1, a)$ não possa ser usado para calcular $x^{(1)}$ pelo referido método. Justifique. **[2.0]**

(b) Fazendo $x^{(0)} = (1, 0, -1)$, mostre que a primeira iterada pode ser calculada resolvendo um sistema linear da forma $A\,w = d$. Obtenha a matriz A e o vector d. **[2.5]**

(c) Se calculasse o vector $w = (w_1, w_2, w_3)$, diga como poderia usá-lo para estimar o erro $||z - x^{(0)}||_2$. Justifique. [Note que não se pede para calcular w]. **[2.5]**

2) Considere os polinómios reais $p(x) = x^4 - x^3 + x^2 - x + 1$ e $r(x)$, sabendo-se que estes polinómios satisfazem as seguintes condições interpolatórias:

$$\begin{array}{lll} p(-2) = r(-2) = 31 & p(-1) = r(-1) = 5 & p(0) = r(0) = 1 \\ p(1) = r(1) = 1 & p(2) = r(2) = 11 & \text{e} \quad r(3) = 30 \ . \end{array}$$

(a) Escreva uma expressão da forma $r(x) = p(x) + c\,\phi(x)$ de modo a relacionar os **[2.0]**

polinómios interpolatórios r e p. Indique a expressão de ϕ e calcule o valor da constante c. Justifique.

(b) Determine o grau de precisão da regra de quadratura **[2.5]**

$$Q(f) = \frac{10}{9}\, f\left(-2\sqrt{\frac{3}{5}}\right) + \frac{16}{9}\, f(0) + \frac{10}{9}\, f\left(2\sqrt{\frac{3}{5}}\right)$$

para aproximar o integral $\int_{-2}^{2} f(x)\, dx$. Justifique.

[2.5] **(c)** Se usasse cálculo exacto diga como poderia aplicar a regra de quadratura $Q(f)$ para obter <u>exactamente</u> o valor de $I = \int_{-2}^{2} r(x)\, dx$. Justifique.

3) Dada a tabela

x	0	1.5	3.0	4.5
$f(x)$	1.00	1.57	2.00	4.30

[2.0] **(a)** Diga o que entende por melhor aproximação de mínimos quadrados da tabela dada, por funções do tipo $g(x) = \alpha\, x + \beta\, sin(x),\ \alpha, \beta \in \mathbb{R}$.

[2.0] **(b)** Determine a matriz A e o vector b de um determinado sistema linear $A\, z = b$, a partir do qual poderia calcular a melhor aproximação referida na alínea anterior (não se pede para resolver o sistema).

[2.0] **(c)** Suponha que $z = [0.87, -0.098]^T$. Qual é o desvio em 1.5? Justifique.

(Teste 23 de Maio 2011, MEEC)

<u>Resolução</u>

1(a) Sejam $x = (x_1, x_2, x_3)$ e $f(x) = (4\, x_1 + x_2^3 + x_3 - 7, x_1\, x_3 + 5\, x_2 - 1, x_1^2 - x_2^2 + x_3^3 + 5)$. Se $f'(x^{(0)})$ for singular não poderá calcular $x^{(1)}$ usando as fórmulas do método. Ora, para $x^{(0)} = (0, 1, a)$, obtém-se

$$f'(x) = \begin{bmatrix} 4 & 3\, x_2^2 & 1 \\ x_3 & 5 & x_1 \\ 2\, x_1 & -2\, x_2 & 3\, x_3^2 \end{bmatrix} \quad \Rightarrow \quad f'(x^{(0)}) = \begin{bmatrix} 4 & 3 & 1 \\ a & 5 & 0 \\ 0 & -2 & 3\, a^2 \end{bmatrix}.$$

Como $det(f'(x^{(0)}) = 4 \times 5 \times 3\, a^2 - a\, (9\, a^2 + 2) = -9\, a^3 + 60\, a^2 - 2\, a = 0$, fazendo por exemplo $a = 0$, resulta $f'(0, 1, 0)$ singular e $x^{(0)} = (0, 1, 0) \in D$.

1(b)

$$\begin{aligned} & x^{(0)} = (1, 0, -1) \\ & A\, w = d \quad \Leftrightarrow \quad J_f(x^{(0)})\, w = -f(x^{(0)}), \end{aligned}$$

donde,

$$A = \begin{bmatrix} 4 & 0 & 1 \\ -1 & 5 & 1 \\ 2 & 0 & 3 \end{bmatrix} \quad \text{e} \quad d = -\begin{bmatrix} -4 \\ -2 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ -5 \end{bmatrix}.$$

1(c) Como $x^{(1)} = x^{(0)} + w$ e $z - x^{(0)} \simeq x^{(1)} - x^{(0)} = w$, resulta

$$||z - x^{(0)}||_2 \simeq \sqrt{w_1^2 + w_2^2 + w_3^2} \ .$$

2(a) Como p e r interpolam os 5 primeiros nós e r interpola mais o ponto $(3, 30)$, tem-se

$$r(x) = p(x) + c\,(x+2)(x+1)\,x(x-1)\,(x-2) \ .$$

Sabe-se que $c = r[-2,-1,0,1,2,3]$, sendo r interpolador dos valores dados para os nós $-2,-1,0,1,2,3$. Como $r(3) = 30$ e $p(3) = 3^4 - 3^3 + 3^2 - 3 + 1 = 61$, obtém-se

$$c = \frac{r(3) - p(3)}{5 \times 4 \times 3 \times 2} = -31/120 \ .$$

Pode verificar que

x_i	r_i	$r[..]$	$r[...]$	$r[....]$	$r[.....]$	$r[......]$
-2	31					
		-26				
-1	5		11			
		-4		-3		
0	1		2		1	
		0		1		$-31/120$
1	1		5		$-7/24$	
		10		$-1/6$		
2	11		9/2			
		19				
3	30					

Logo, $r(x) = p(x) - 31/120(x+2)\,(x+1)\,x(x-1)\,(x-2)$, onde

$$p(x) = 31 - 26\,(x+2) + 11\,(x+2)\,(x+1) - 3\,(x+2)\,(x+1)\,x + (x+2)\,(x+1)\,x\,(x-1) \ .$$

2(b) Atendendo a que

$$\begin{array}{l} Q(1) = 36/9 = 4 = \int_{-2}^{2} dx = I(1) \\ Q(x) = I(x) = 0 \\ Q(x^2) = 2 \times (10/9) \times 2^2 \times 3/5 = 2^4/3 = I(x^2) \\ Q(x^3) = 0 = I(x^3) \\ Q(x^4) = 2 \times (10/9) \times 2^2 \times (3^2/5^2) = 2^6/5 = I(x^4) \\ Q(x^5) = 0 = I(x^5) \\ Q(x^6) = 2 \times (10/9) \times 2^6 \times (3^3/5^3) = 2^8/5^2 \quad \text{e} \\ I(x^6) = 2^8/7 \neq Q(x^6) \ . \end{array}$$

Conclui-se que a regra Q é de grau 5, porquanto as relações anteriores implicam que a regra é exacta para qualquer polinómio de grau ≤ 5 mas não é exacta para o monómio x^6. Tal equivale a dizer-se que a regra possui grau 5 de precisão.

2(c) Como r é polinómio de grau 5 a regra é exacta para r. Assim,

$$I = 10/9\, r(-2\sqrt{3/5}) + 16/9\, r(0) + 10/9\, r(2\sqrt{3/5}) \ .$$

3(a) A melhor aproximação de mínimos quadrados da tabela é a função $\tilde{g}(x) = \tilde{\alpha}\, x + \tilde{\beta}\, \sin(x)$, tal que é mínimo o valor de $||f-g||^2 = \sum_{i=0}^{3}(f(x_i) - \alpha\, x_i - \beta\, \sin(x_i))^2$, ou seja, $||f-\tilde{g}||^2 \leq ||f-g||^2$, para quaisquer valores reais de α e β.

3(b) Fazendo

$f = (1.00, 1.57, 2.00, 4.30)^T$
$\phi_0 = (0, 1.5, 3.0, 4.5)^T$
$\phi_1 = (\sin(0), \sin(1.5), \sin(3.0), \sin(4.5))^T = (0, 0.997495, 0.14112, -0.977530)^T$
$z = (\alpha, \beta)^T$,

obtém-se o sistema de equações normais $A\, z = b$, com

$$A \simeq \begin{bmatrix} 31.5 & -2.47928 \\ -2.47928 & 1.97048 \end{bmatrix} \qquad b \simeq \begin{bmatrix} 27.705 \\ -2.35507 \end{bmatrix} \ .$$

3(c) Sendo $\tilde{g}(x) = 0.87\, x - 0.098\, \sin(x)$, resulta $\tilde{g}(1.5) = 1.20725$, pelo que o desvio pretendido é $d = f(1.5) - \tilde{g}(1.5) \simeq 0.36$.

A.2.5

Parte I

1) Considere a função real $f(x) = 1/(x-a), \quad x \neq a$.

[2.0] **(a)** Sejam $x = 0.12345 \times 10^{-5}$ e $a = 0.12340 \times 10^{-5}$. Calcule exactamente o erro relativo (expresso em percentagem) que se comete ao calcular $f(x)$ num sistema decimal de ponto flutuante com 4 dígitos na mantissa e arredondamento simétrico.

[2.0] **(b)** Diga, justificando, se a função f considerada é bem condicionada para valores de x próximos do valor de a dado. Sugestão: poderá justificar a sua resposta levando em consideração, nomeadamente, o resultado que obteve na alínea anterior.

2) Dado um número real $a > 0$, pretende-se calcular aproximações de $\sqrt{a}$ mediante aplicação de um processo iterativo.

[2.0] **(a)** Mostre que se aplicar o método de Newton obtém uma sucessão de iteradas da forma $x_{n+1} = \frac{1}{2}\,(x_n + \frac{a}{x_n})$.

[2.5] **(b)** Prove que é válida a igualdade

$$x_{n+1}^2 - a = \left(\frac{x_n^2 - a}{2\, x_n}\right)^2 \ .$$

Admitindo que a sucessão $(x_n)_{n\geq 0}$ converge para $\sqrt{a}$, mostre que a sucessão converge quadraticamente. Justifique.

[2.5] **(c)** Sendo a o seu número de aluno, calcule uma aproximação de $\sqrt{a}$, com erro absoluto inferior a 10^{-2}, usando o referido método. Justifique a escolha que fizer da aproximação inicial x_0. Sugestão: comece por determinar um número natural N, tal que $N < \sqrt{a} < N+1$.

3) Considere o sistema linear $A\,x = b$, sendo

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1/5 & 1/4 & 4/5 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 1 \\ 17/20 \end{bmatrix} .$$

a) Escreva fórmulas iterativas que considere adequadas para obter aproximações da solução x do sistema. **[2.0]**

b) A partir das fórmulas que considerou na alínea anterior, obtenha a matriz de iteração do respectivo método. Diga, justificando, se uma vez escolhida a aproximação inicial $x^{(0)} = (0,0,0)^T$, o método é convergente para a solução do sistema, independentemente da norma vectorial que usar. Justifique. **[2.0]**

c) Partindo da aproximação $x^{(0)} = (-1,0,1)^T$, obtenha a iterada $x^{(3)}$ e calcule o valor exacto de $||x - x^{(3)}||_\infty$. **[2.0]**

4) Considere o método iterativo $R\,x^{(k+1)} = S\,x^{(k)} + c,\ k = 0,1,\ldots$, aplicado à resolução de um sistema linear $M\,x = c$, onde M é matriz não singular e c é um vector coluna arbitrário. Sabe-se que as entradas da matriz M são: $m_{i,i} = 1$, $m_{i,j} = 1/(i+j-1)$ (se $i \neq j$), para $i,j = 1,2,3$. Além disso a matriz R é diagonal, de entradas $r_{i,i} = i+1$, para $i = 1,2,3$. Obtenha a matriz S, e prove que o método converge para a solução x, independentemente da aproximação inicial que escolher. **[3.0]**

Parte II

1) Dado o sistema de equações **[2.0]**

$$\begin{cases} 2\,x_1^2 + 2\,x_1 + 3\,x_3 &= 1 \\ e^{x_1} - 4 &= 0 \\ x_1 + x_2 = 0 \end{cases}$$

Obtenha a primeira iterada do método de Newton, tomando para aproximação inicial da solução o vector $x^{(0)} = (1,-1,-2)^T$. Apresente todos os cálculos que efectuar, dando os resultados arredondados para 6 dígitos decimais.

2) Considere a função real $y(t) = \dfrac{A}{2}\,t^2 + B\,t + C$ da qual se conhecem os valores a seguir tabelados

t	0.2	0.3	0.4	0.5
$y(t)$	0.940	0.655	0.577	0.706

(a) Aplique o método de interpolação de Newton para determinar os valores de A, B e C. Justifique a escolha que fizer dos nós de interpolação. **[2.0]**

(b) Sendo s um ponto arbitrário do intervalo $[0.2, 0.5]$, qual é o valor máximo do erro absoluto de interpolação que comete, ao calcular um valor aproximado de $y(s)$ por interpolação linear (isto é, com polinómios de grau não superior a 1)? Justifique. **[2.5]**

(c) Se utilizasse todos os pontos tabelados para determinar os parâmetros A, B e C mediante aplicação do método dos mínimos quadrados, obteria o mesmo resultado que na alínea (a)? Justifique. **[2.0]**

[2.0] **(d)** Obtenha uma fórmula de quadratura $Q(f) = 1/30\, f(3/10) + \beta\, f(2/5) + \gamma\, f(1/2)$, que lhe permita aproximar o valor $\displaystyle\int_{3/10}^{1/2} f(x)\, dx$, aplicando o método dos coeficientes indeterminados, de modo que ela seja pelo menos de grau 1 de precisão.

[2.0] **(e)** A regra Q anterior é uma regra de Newton-Cotes fechada. Qual é a sua designação habitual? Justifique. [Caso não tenha resolvido a alínea (d), faça $\beta = 2/15$ e $\gamma = 1/30$].

[2.5] **(f)** Calcule exactamente o valor de $\displaystyle\int_{0.3}^{0.5} y(t)\, dt$ aplicando uma regra de quadratura que considere adequada para esse efeito. Justifique a escolha que fizer dessa regra.

3. Considere o problema de valor inicial

$$\begin{aligned} y' &= x + e^y \\ y(1) &= 0\ . \end{aligned}$$

Sabe-se que $y(1.2) = 0.472266\ldots$.

[2.5] **(a)** Obtenha uma aproximação de $y(1.2)$ aplicando o método de Euler com passo $h = 0.1$. Diga, justificando, quantos algarismos significativos possui o resultado que calculou.

[2.5] **(b)** Para a equação diferencial dada, deduza a fórmula do respectivo método de Taylor de segunda ordem.

(Exame 11 Junho 2011, MEEC)

Resolução (Parte I)

1(a) Seja $z = 1/(x-a) = 1/(5\times 10^{-5}\times 10^{-5}) = 10^{10}/5$. Como o valor de a dado tem representação exacta no sistema $FP(10,4)$, o resultado do cálculo de $f(x)$ será

$$\bar{z} = fl(\frac{1}{fl(x)-a})\ .$$

Ora, $fl(x) - a = (0.1235 - 0.1234)\times 10^{-5} = 10^{-9}$. Então, $\bar{z} = fl(10^9) = 0.1000\times 10^{10}$. Assim, atendendo a que $z - \bar{z} = 10^{10}/5 - 10^9 = 10^9$, resulta

$$\delta_{\bar{z}} = \frac{z-\bar{z}}{z} = \frac{1}{2} = 0.5 = 50\,\%\ .$$

1(b) Atendendo a que o erro relativo que se comete na passagem do valor x dado a $fl(x)$ (única perturbação existente no cálculo de $f(x)$), é tal que $|\delta_{fl(x)}| \leq 0.5\times 10^{-3} = 0.05\,\%$, visto que o erro propagado à função é muito maior ($50\,\%$), podemos concluir que a função é mal condicionada para valores de x próximos do valor de a dado.

2(a) Pretende-se determinar um número real $x > 0$ tal que $x^2 = a$. Para $f(x) = x^2 - a = 0$, as iteradas do método resultam de $x_{n+1} = x_n - f(x_n)/f'(x_n)$, isto é,

$$x_{n+1} = x_n \frac{x_n^2 - a}{2\,x_n} = \frac{1}{2}\,(x_n + \frac{a}{x_n}), \quad n = 0, 1, \ldots \ .$$

2(b) De $x_{n+1}^2 = \frac{1}{4}\,(x_n^2 + 2\,a + \frac{a^2}{x_n^2})$, resulta

$$x_{n+1}^2 - a = \frac{1}{4}\,\frac{x_n^4 - 2\,a\,x_n^2 + a^2}{x_n^2} = (\frac{x_n^2 - a}{2\,x_n})^2.$$

Ou seja,

$$\frac{x_{n+1} - \sqrt{a}}{(x_n - \sqrt{a})^2} = \frac{(x_n + \sqrt{a})^2}{4\,x_n^2(x_{n+1} + \sqrt{a})}, \quad x_n \neq 0 \ .$$

Como por hipótese $(x_n)_{n\geq 0}$ converge para $\sqrt{a}$, passando ao limite obtém-se:

$$\lim_{n\to\infty} \frac{x_{n+1} - \sqrt{a}}{(x_n - \sqrt{a})^2} = \frac{(2\,\sqrt{a})^2}{4\,a \times 2\,\sqrt{a}} = \frac{1}{2\,\sqrt{a}} \neq 0,$$

significando que a convergência é quadrática.

2(c) O menor número de aluno do curso é $a > 40\,000$, logo

$$N < \sqrt{a} < N + 1, \quad \text{com} \quad N > 200 \ .$$

Basta uma iteração do método para se obter uma aproximação de $\sqrt{a}$ com erro inferior a 10^{-2}.

Nota

Com efeito, para $x_0 = N + 1$, atendendo a que $f''(x) \times f(x_0) > 0 \quad \forall x \geq \sqrt{a}$, sabemos que o método converge para $\sqrt{a}$ e a convergência é monótona. Além disso, atendendo à fórmula de erro do método,

$$x_{n+1} - \sqrt{a} = -\frac{1}{2\,x_n}(x_n - \sqrt{a})^2, \quad \forall n \geq 0 \ .$$

Assim, visto que $x_n > N$ e $x_0 - \sqrt{a} < 1$, são válidas as desigualdades

$$|x_1 - \sqrt{a}| < \frac{1}{2\,N}\,(x_0 - \sqrt{a})^2 < \frac{1}{2\,N} \ .$$

Como $1/(2\,N) < 1/400 < 10^{-2}$, basta uma iteração do método para se obter uma aproximação de $\sqrt{a}$ com erro inferior a 10^{-2} .

3(a) O sistema é equivalente a

$$\begin{cases} x_1 = -\frac{17}{4} + \frac{5}{4}x_2 + 4\,x_3 \\ x_2 = 1 \\ x_3 = 1 \end{cases} \qquad k = 0, 1, \ldots$$

donde resultam as fórmulas iterativas

$$\begin{array}{ll} x_1^{(k+1)} = -\dfrac{17}{4} + \dfrac{5}{4} x_2^{(k)} + 4\, x_3^{(k)} & \\ x_2^{(k+1)} = 1 & k = 0, 1, \ldots \\ x_3^{(k+1)} = 1 & \end{array}$$

da forma $x^{(k+1)} = C\, x^{(k)} + d$.

3(b) Da alínea anterior resulta imediatamente

$$C = \begin{bmatrix} 0 & 5/4 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .$$

O polinómio característico respectivo é $p(\lambda) = \lambda^3$, logo $\rho(C) = 0$, pelo que o método converge (usando uma qualquer norma) independentemente da aproximação inicial escolhida.

3(c)

$$\begin{array}{l} x^{(1)} = (-17/4 + 4, 1, 1)^T = (-1/4, 1, 1)^T \\ x^{(2)} = (-17/4 + 5/4 + 4, 1, 1)^T = (1, 1, 1)^T = x \\ x^{(3)} = x \ . \end{array}$$

Logo, $||x - x^{(3)}||_\infty = 0$.

4) Como $M\, x = c \ \Leftrightarrow (R - S)\, x = c \ \Leftrightarrow R\, x = S\, x + c$, tem-se:

$$M = R - S = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1 & 1/4 \\ 1/3 & 1/4 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} - \begin{bmatrix} 1 & -1/2 & -1/3 \\ -1/2 & 2 & -1/4 \\ -1/3 & -1/4 & 3 \end{bmatrix} .$$

Por conseguinte, a matriz de iteração do método em causa é:

$$C = R^{-1}\, S = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/4 \end{bmatrix} \begin{bmatrix} 1 & -1/2 & -1/3 \\ -1/2 & 2 & -1/4 \\ -1/3 & -1/4 & 3 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/4 & -1/6 \\ -1/6 & 2/3 & -1/12 \\ -1/12 & -1/16 & 3/4 \end{bmatrix} .$$

Atendendo a que

$$||C||_\infty = max(11/12, 11/12, 43/48) = 11/12 < 1,$$

o método converge $\forall\, x^{(0)}$ de partida.

Resolução (Parte II)

1(a) Sendo $f(x_1, x_2, x_3) = (2\, x_1^2 + 2\, x_1 + 3\, x_3 - 1, e^{x_1} - 4, x_1 + x_2)^T$, obtém-se

$$J_f(x_1, x_2, x_3) = \begin{bmatrix} 4\, x_1 + 2 & 0 & 3 \\ e^{x_1} & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} .$$

Para $x^{(0)} = (1, -1, -2)^T$, vem $f(x^{(0)}) = (-3, e-4, 0)^T$. A primeira iterada do método obtém-se resolvendo o sistema linear

$$J_f(x^{(0)})\,\Delta x^{(0)} = -\,f(x^{(0)})), \qquad \text{e} \quad x^{(1)} = x^{(0)} + \Delta x^{(0)}$$

$$\begin{bmatrix} 6 & 0 & 3 \\ e & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta\, x_1 \\ \Delta\, x_2 \\ \Delta\, x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 4-e \\ 0 \end{bmatrix} .$$

Assim,

$$\begin{aligned} \Delta\, x_1 &= \frac{4-e}{e} \simeq 0.471518 \\ \Delta\, x_2 &= -\Delta\, x_1 \simeq -0.471518 \\ \Delta\, x_3 &= \frac{3 - 6\,\Delta\, x_1}{3} \simeq 0.0569645 \ . \end{aligned}$$

Por conseguinte $x^{(1)} \simeq (1.47152, -1.47152, -1.94304)^T$.

2(a) Como a função dada $y(t)$ é polinómio de grau 2, atendendo a que o polinómio interpolador para quaisquer 3 nós distintos é único, podemos usar quaisquer 3 pontos tabelados para determinar o respectivo polinómio interpolador, o qual é idêntico à função dada. A partir da tabela de diferenças divididas

t_i	y_i	$y[..]$	$y[...]$
0.2	0.94		
		-2.85	
0.3	0.655		10.35
		-0.78	
0.4	0.577		

obtém-se:

$$p(t) = 0.94 - 2.85\,(t-0.2) + 10.35\,(t-0.2)\,(t-0.3) = 2.131 - 8.025\,t + 10.35\,t^2 \ .$$

Assim, $A/2 = 10.35 \Leftrightarrow A = 20.70$, $B = -8.025$ e $C = 2.131$.

2(b) Sejam $t_0 = 0.2, t_1 = 0.3$ e $t_2 = 0.4$. Sabe-se que existe $\mu \in (t_i, t_{i+1})$ tal que

$$y(s) - p(s) = \frac{y''(\mu)}{2}\,(s-t_i)\,(s-t_{i+1}) = \frac{A}{2}\,(s-t_i)\,(s-t_{i+1}),$$

onde p é polinómio interpolador nos nós consecutivos t_i e t_{i+1}. Assim,

$$M = max_{0.2\le s\le 0.5}|y(s) - p(s)| = A/2\ max_{0.2\le s\le 0.5}\,|(s-t_i)\,(s-t_{i+1})| \ .$$

Ora, o polinómio $w(s) = (s-t_i)\,(s-t_{i+1}) = s^2 - (t_i + t_{i+1})\,t + t_i\,t_{i+1}$ possui extremo no ponto $\tilde{s} = (t_i + t_{i+1})/2$, de valor $w(\tilde{s}) = (t_{i+1} - t_i)/2 \times (t_i - t_{i+1})/2 = -0.1^2/4$. Logo,

$$M = A/8 \times 0.1^2 = 0.025875 \ .$$

2(c) Uma vez que a melhor aproximação polinomial, de grau ≤ 2, de mínimos quadrados, é única, atendendo a que $p(0.5) = 0.706 = y(0.5)$, conclui-se que os desvios de

p em todos os valores tabelados são nulos. Por conseguinte p coincide com a melhor aproximação de mínimos quadrados pretendida, pelo que o resultado seria o mesmo que na alínea (a).

2(d) A regra é exacta para qualquer polinómio de grau ≤ 1 se e só se é exacta para os monómios 1 e x, isto é,

$$\begin{cases} \beta + \gamma & = \int_{3/10}^{1/2} dx - 1/30 = 1/2 - 3/10 - 1/30 = 1/6 \\ 2/5\,\beta + 1/2\,\gamma & = \int_{3/10}^{1/2} x\,dx - 3/300 = 2/25 - 3/300 = 7/100\ . \end{cases}$$

Logo,

$$\beta = \frac{\begin{vmatrix} 1/6 & 1 \\ 7/100 & 1/2 \end{vmatrix}}{1/10} = 10\,(1/12 - 7/100) = 2/15,$$

$$\gamma = \frac{\begin{vmatrix} 1/ & 1/6 \\ 2/5 & 7/100 \end{vmatrix}}{1/10} = 10\,(7/100 - 1/15) = 1/30\ .$$

Assim,

$$Q(f) = 1/30\,f(3/10) + 2/15\,f(2/5) + 1/30\,f(1/2)\ .$$

2(e) Sendo $h = (1/2 - 3/10)/2 = 2/20 = 1/10$, a regra de Simpson (que é de grau 3), escreve-se:

$$\begin{aligned} S(f) &= 1/30\,[f(3/10) + 4\,f(2/5) + f(1/2)] \\ &= 1/30 f(3/10) + 2/15 f(2/5) + 1/30\,f(1/2) = Q(f)\ . \end{aligned}$$

2(f) Dado que $y(t)$ é polinómio de grau 2, a regra de Simpson (ou seja, a regra Q), é exacta quando aplicada a y, isto é,

$$\begin{aligned} \textstyle\int_{0.3}^{0.5} y(t)\,dt &= 1/30\,[y(0.3) + 4\,y(0.4) + y(0.5)] = 1/30\,(0.655 + 4 \times 0.577 + 0.706) \\ &= 0.1223\ . \end{aligned}$$

3(a) Para $f(x, y) = x + e^y$, $h = 0.1$, $x_0 = 1$, $y_0 = 0$, obtém-se:

$$\begin{aligned} y_1 &\simeq y(1.1) = y_0 + h\,f(x_0, y_0) = 0 + 0.1\,(1 + e^0) = 0.2 \\ y_2 &\simeq y(1.2) = y_1 + h\,f(x_1, y_1) = 0.2 + 0.1\,(1.1 + e^{0.2}) = 0.43214\ . \end{aligned}$$

Como $y(1.2) = 0.472266\ldots$, conclui-se imediatamente que o valor calculado de y_2 possui 1 algarismo significativo, visto o seu erro absoluto ser aproximadamente 0.04 unidades.

3(b) Para o passo h, atendendo que que o desenvolvimento de Taylor de segunda ordem da função y é

$$y(x + h) = y(x) + h\,f(x, y) + h^2/2\,y''(x) + h^3/3!\,y^{(3)}(\xi), \quad \xi \in (x, x + h),$$

como $y'(x) = f(x, y) = x + e^y$, obtém-se

$$y''(x) = 1 + e^y \times (x + e^y)\ .$$

Assim, a fórmula do método em causa escreve-se:

$$y_0 = 0$$
$$y_{i+1} = y_i + h\,(x_i + e^{y_i}) + h^2/2\,(1 + e^{y_i} \times (x_i + e^{y_i})), \qquad i = 0, 1, \ldots\ .$$

A.2.6

Observação: O símbolo α em algumas questões designa o último dígito do seu número de aluno.

1) Considere o sistema

$$\left\{ \begin{array}{lll} 3\,x_1 + x_2 & = 4 & \\ \sin(x_1) - 2\,x_2 & = 1 & (*) \\ x_3 & = 1 & \end{array} \right.$$

onde $x = (x_1, x_2, x_3) \in \mathbb{R}^3$.

(a) Fazendo $w^{(0)} = [0, 1, \alpha]^T$, mostre que a primeira iterada $w^{(1)}$ do método de Newton aplicado ao sistema $(*)$ pode ser calculada resolvendo um sistema linear da forma $A\,w = c$, onde **[1.5]**

$$A = \left[\begin{array}{ccc} 3 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & 1 \end{array} \right] \quad \text{e} \quad c = [3, 3, 1 - \alpha]^T.$$

Calcule exactamente $||w - w^{(1)}||_1$.

(b) Diga, justificando, se poderá aplicar o método iterativo de Jacobi para aproximar a solução w do sistema linear $A\,w = c$. **[1.5]**

(c) Nesta alínea suponha que $\alpha = 1$. Partindo de $w^{(0)} = [1, 1, 1]^T$, calcule a segunda iterada $w^{(2)}$ do método de Gauss-Seidel, bem como um majorante para $||w - w^{(2)}||_1$. **[1.5]**

2) Considere a função real

$$f(x) = \left\{ \begin{array}{l} \alpha + 1 + 4\cos x, \quad \text{se} \quad 0 \leq x < \pi \\ \dfrac{x^2}{2} - x + 1, \quad \text{se} \quad x \geq \pi, \end{array} \right.$$

onde α tem o significado referido na Observação.

Comece por determinar uma tabela de valores $(x_i, f(x_i))$, onde $f(x_i)$ ou é exacto ou possui pelo menos 5 algarismos significativos, sendo $x_i = 3, 4, 5, 6$.

(a) Usando o polinómio $q(x)$, interpolador da função nos três últimos pontos tabelados, obtenha o valor $q(5.2)$. Calcule o respectivo erro de interpolação. Justifique. **[1.5]**

(b) Mediante funções aproximantes do tipo **[1.5]**

$$\Psi(x) = c_1\, sin(x) + c_2\, sin(2\,x), \quad c_1, c_2\ \in \mathbb{R}$$

obtenha a matriz A de um sistema linear $A\,c = \omega$ cuja solução lhe permite obter a melhor aproximação de mínimos quadrados dos três primeiros pontos tabelados. Apresente a matriz pedida A, cujas entradas estejam arredondadas na forma $\pm d_1.d_2$ (por exemplo, 1.5). Note que não é necessário calcular a solução do sistema referido mas deverá indicar as componentes de c e ω.

[1.5] **(c)** Diga, justificando, se poderá aplicar a regra de Simpson simples para aproximar $\int_3^5 f(x)dx$. No caso afirmativo, como estimaria o respectivo erro?

[1.0] **(d)** Aplique a regra dos trapézios composta , com passo $h = 1$, para aproximar

$$\int_4^6 (\alpha + 1)\, f(x)dx \ .$$

(Teste 21 Dezembro 2012, MEEC)

Resolução

1(a) Para $f(x_1, x_2, x_3) = (3\,x_1 + x_2 - 4, \sin x_1 - 2\,x_2 - 1, x_3 - 1)^T$, resulta $f(w^{(0)}) = (-3, -3, \alpha - 1)^T$. Como

$$J(x^{(0)}) = \begin{bmatrix} 3 & 1 & 0 \\ \cos(x_1) & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

a primeira iterada do método de Newton é calculada resolvendo o sistema linear $J(w^{(0)})\,\Delta w^{(0)} = -f(w^{(0)})$, ou seja, o sistema $A\,w = c$ dado.

Como $\Delta w^{(0)} = w^{(1)} - w^{(0)} = w$, logo $||w - w^{(1)}||_1 = ||w^{(0)}||_1 = 1 + \alpha$.

1(b) A partir do sistema $A\,w = c$, obtém-se

$$\begin{cases} w_1 = \dfrac{3 - w_2}{3} \\ w_2 = \dfrac{-3 + w_1}{2} = \dfrac{-3 + \frac{3 - w_2}{3}}{2} = \dfrac{-6 - w_2}{6} \\ w_3 = 1 - \alpha \end{cases} \qquad (**)$$

pelo que o metodo iterativo de Jacobi tem a forma

$$\begin{cases} w_1^{(k+1)} = \dfrac{3 - w_2^{(k)}}{3} \\ w_2^{(k+1)} = \dfrac{-3 + w_1^{(k)}}{2} \\ w_3^{(k+1)} = 1 - \alpha \end{cases} , \quad k = 0, 1, \ldots \ .$$

Logo, a respectiva matriz de iteração é

$$C_J = \begin{bmatrix} 0 & -1/3 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .$$

Dado que $||C_J||_\infty = max(1/2, 1/3) = 1/2 < 1$, este método converge para a solução do sistema dado, independentemente do vector inicial $w^{(0)}$.

1(c) Das equações $(**)$ resulta para o método de Gauss-Seidel,

$$\begin{cases} w_1^{(k+1)} = 1 - w_2^{(k)}/3 \\ w_2^{(k+1)} = -1 - w_2^{(k)}/6 \\ w_3^{(k+1)} = 1 - \alpha \end{cases} , \quad k = 0, 1, \ldots$$

i.e.,

$$w^{(k+1)} = \begin{bmatrix} 0 & -1/3 & 0 \\ 0 & -1/6 & 0 \\ 0 & 0 & 0 \end{bmatrix} w^{(k)} + \begin{bmatrix} 1 \\ -1 \\ 1-\alpha \end{bmatrix} .$$

Por conseguinte, $||C_{GS}||_1 = max(0, 1/2) = 1/2$. Assim, para $\alpha = 1$, obtém-se

$$\begin{aligned} w^{(1)} &= [1 - 1/3, -1 - 1/6, 0]^T = [2/3, -7/6, 0]^T \\ w^{(2)} &= [1 + 7/18, -1 + 7/36, 0]^T = [25/18, -29/36, 0]^T, \end{aligned}$$

logo $w^{(2)} - w^{(1)} = [13/18, 13/36, 0]^T$, e

$$\begin{aligned} ||w - w^{(2)}|| &\leq \frac{||C_{GS}||_1}{1 - ||C_{GS}||_1} ||w^{(2)} - w^{(1)}||_1 \\ &\leq ||w^{(2)} - w^{(1)}||_1 = 13/18 + 13/36 = 13/12 \simeq 1.08333 . \end{aligned}$$

2(a)

x_i	f_i	$f[..]$	$f[...]$
3	$\alpha + 1 + 4\cos 3$		
4	5		
		7/2	
5	17/2		1/2
		9/2	
6	13		

Para $x \geq \pi$ a função f é polinomial de grau 2. Por conseguinte o polinómio interpolador nos três últimos nós da tabela coincide com f. Isto é, $q(x) = f(x)$, $x \geq \pi$. Logo, $q(5.2) = f(5.2) = 9.32$. De facto, da tabela de diferenças divididas acima resulta

$$q(x) = 5 + 7/2\,(x-4) + 1/2\,(x-4)\,(x-5) = x^2/2 - x + 1 .$$

2(b) Sendo $f = (\alpha + 1 + 4\cos(3), 5, 17/2)^T$ e

$$\begin{aligned} \phi_0 &= (\sin 3, \sin 4, \sin 5)^T \\ \phi_1 &= (\sin 6, \sin 8\ \sin 10)^T \\ \Psi &= c_1\,\phi_0 + c_2\,\phi_1 \end{aligned}$$

a melhor aproximação de mínimos quadrados satisfaz a condição $||f - \Psi||_2^2 = min \quad \forall c_1, c_2 \in \mathbb{R}$ se e só se

$$\begin{bmatrix} (\phi_0, \phi_0) & (\phi_0, \phi_1) \\ (\phi_0, \phi_1) & (\phi_1, \phi_1) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} (f, \phi_0) \\ (f, \phi_1) \end{bmatrix} .$$

Ora,

$$(\phi_0,\phi_0) = \sin^2(3)+\sin^2(4)+\sin^2(5) \simeq 1.5122$$
$$(\phi_0,\phi_1) = \sin(3)\,\sin(6)+\sin(4)\,\sin(8)+\sin(5)\,\sin(10) \simeq -0.266505$$
$$(\phi_1,\phi_1) = \sin^2(6)+\sin^2(8)+\sin^2(10) \simeq 1.35286\ .$$

Logo, a matriz A do sistema a resolver, arredondada, é

$$A = \begin{bmatrix} 1.5 & -0.27 \\ -0.27 & 1.4 \end{bmatrix}$$

e $c = (c_1, c_2)^T$, $\omega = ((\phi_0, f), (\phi_1, f))^T$.

2(c) A função f não é contínua em $x = \pi$. Assim, embora a regra $S(f) = 1/3\,(f(3) + 4\,f(4) + f(5))$ produza um número real, a fórmula de erro não é aplicável pois esta só é válida para funções de classe C^4 (pelo menos), no intervalo considerado.

2(d) Seja $I = \displaystyle\int_4^6 (\alpha+1)\,f(x)dx = (\alpha+1)\int_4^6 f(x)dx$. Pela regra dos trapézios

$$\int_4^6 f(x)dx \simeq h/2\,(f(4)+2\,f(5)+f(16)) = 1/2(5+17+13) = 35/2 = 17.5\ .$$

Assim, $I \simeq (\alpha+1) * 17.5$.

A.2.7

1) Sabe-se que os números $\tilde{a} = 3.1415$ e $\tilde{b} = -3.1425$ resultaram de arredondamentos simétricos para 5 dígitos decimais.

[1.5] **(a)** Estime o erro absoluto do valor de $\tilde{y} = \dfrac{\tan(\tilde{a}+\tilde{b})}{2}$. Apresente todos os cálculos que efectuar.

[1.0] **(b)** Quantos algarismos significativos pode garantir para o valor de $\tilde{y}$ referido na alínea anterior? Justifique.

2) Considere a função geradora $g(x) = \sin(\alpha\,x)$, onde $\alpha > 0$ é um parâmetro real, e o processo iterativo

$$x_0 = 1 \qquad x_{n+1} = g(x_n), \quad n = 0, 1, \ldots \qquad (*)$$

Sabe-se que para $1.5 \leq \alpha \leq 2$, a função g possui um único ponto fixo z no intervalo $[0.9, 1]$. Nas alíneas (a), (b) e (c) a seguir admita que $\alpha = 2$.

[1.0] **(a)** Mostre que o ponto fixo z é raiz da equação $\sin(x)\,\cos(x) - \dfrac{x}{2} = 0$.

[1.0] **(b)** Verifique que no intervalo considerado estão satisfeitas as condições de convergência do método de Newton quando aplicado ao cálculo de aproximações de z .

[1.5] **(c)** Obtenha uma aproximação de z, com erro absoluto inferior a 10^{-5}, escolhendo x_0

de modo que a sucessão de iteradas do método de Newton seja monótona. Justifique.

[1.5] **(d)** Se fixar $0 < \alpha < 1$, poderá afirmar que a sucessão $(*)$ é convergente? Em caso afirmativo diga, justificando, se a convergência é supralinear.

3) Dado o sistema linear $A\,x = b$, com

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 10^{-4} \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 3 \\ 10^{-4} \end{bmatrix},$$

considere o sistema linear $A\,u = c$, onde $c = b + [0, -2 \times 10^{-4}]^T$.

(a) Obtenha $||x - u||_\infty / ||x||_\infty$. **[1.0]**

(b) Calcule $cond_\infty(A)$. Diga, justificando, se pode afirmar que o sistema $A\,x = b$ é bem condicionado. **[1.5]**

(Teste 8 Abril 2013, MEC/LEGM)

Resolução

1 (a) Sabemos que para as aproximações $\tilde{a} = 3.1415$ e $\tilde{b} = -3.1425$, o respectivo erro absoluto satisfaz a condição $|e_{\tilde{a}}| \leq 0.5 \times 10^{-4}$, $|e_{\tilde{b}}| \leq 0.5 \times 10^{-4}$. Da fórmula de propagação do erro absoluto obtém-se

$$e_{\tilde{y}} \simeq \frac{1}{2}\sec^2(\tilde{a} + \tilde{b})\, e_{\tilde{a}} + \frac{1}{2}\sec^2(\tilde{a} + \tilde{b})\, e_{\tilde{b}},$$

e por conseguinte o erro absoluto de $\bar{y}$ pode ser majorado por

$$\begin{aligned} |e_{\tilde{y}}| &\leq \sec^2(\tilde{a} + \tilde{b}) \times 0.5 \times 10^{-4} \simeq 1.0 \times 0.5 \times 10^{-4} \\ &\leq 0.5 \times 10^{-4} \ . \end{aligned}$$

1 (b) Como

$$y = \frac{\tan(3.141\cdots - 3.142\cdots)}{2} = -0.00050\cdots = -0.50\cdots \times 10^{-3},$$

e atendendo à alínea anterior, tem-se

$$|e_{\tilde{y}}| \leq 0.5 \times 10^{-4} = 0.5 \times 10^{-3-1} \ .$$

Assim, podemos garantir que a aproximação

$$\tilde{y} = \tan(\bar{a} + \bar{b})/2 = -0.00050\cdots = -0.50\cdots \times 10^{-3}$$

possui um algarismo significativo.

2 (a) Um ponto fixo de g é solução da equação $x = \sin(2\,x)$, ou seja,
$x = 2\,\sin(x)\,\cos(x) \Leftrightarrow \sin(x)\,\cos(x) - x/2 = 0$.

2 (b) Seja $I = [0.9, 1]$ e $f(x) = \sin(x)\,\cos(x) - x/2 \in C^2(I)$. Como
(i) $f(0.9) \times f(1) \simeq 0.037 \times (-0.045) < 0$, existe pelo menos um zero de f em I .

Atendendo a que

$$f'(x) = \cos^2(x) - \sin^2(x) - 1/2$$
$$f^{(2)}(x) = -4\,\sin(x)\,\cos(x) < 0 \quad \forall\, x \in I,$$

conclui-se que f' é função estritamente decrescente. Ora, $f'(0.9) < 0$, logo
(ii) $f'(x) < 0\ \forall x \in I$, pelo que existe um só zero de f em I (que é o ponto fixo z da função iteradora g considerada). Além disso,
(iii) $\left|\dfrac{f(0.9)}{f'(0.9)}\right| \simeq |-0.05| < 0.1$ e $\left|\dfrac{f(1)}{f'(1)}\right| |\simeq -0.05| < 0.1$.
Por conseguinte podemos garantir convergência (quadrática) do método de Newton para o ponto z.

2 (c) Como $f^{(2)}$ no intervalo em causa possui o sinal de $f(1) \simeq -0.045$, escolhendo $x_0 = 1$, a convergência do método é monótona:

$$x_1 = x_0 - f(x_0)/f'(x_0) \simeq 0.9504977971$$
$$x_2 = x_1 - f(x_1)/f'(x_1) \simeq 0.947755823\ .$$

Ora, $e_{x_2} \simeq -f(x_2)/f'(x_2) \simeq -8.7 \times 10^{-6} < 10^{-5}$, pelo que $z \simeq 0.9477558$, com erro absoluto inferior a 10^{-5} .

2 (d) Para $0 < \alpha < 1$ e $0 \leq x \leq 1$, tem-se que $\sin(\alpha x) < x$. Assim, para $x_0 = 1$, resulta

$$0 < x_1 = g(x_0) = \sin(\alpha \times 1) < x_0$$
$$0 < x_2 = g(x_1) = \sin(\alpha \times x_1) < x_1$$
$$\vdots$$

donde se conclui por indução que a sucessão de iteradas é constituída por termos positivos e decrescentes. Logo a sucessão tende para o ponto $x = 0$. Ora, $g(0)) = 0$ pelo que $z = 0$ é ponto fixo de g. Como $g'(x) = \alpha\,\cos(\alpha\, x)$, resulta $0 < g'(0) = \alpha < 1$, donde se conclui que a convergência da sucessão é linear.

3 (a) Como $x = [1, 1]^T$ e $u = [2, -1]^T$, tem-se $||x-u||_\infty = max(2, 1) = 2$ e $||x||_\infty = 1$. Assim, $||x - u||_\infty / ||x||_\infty = 2$.

3 (b) Como

$$A^{-1} = \frac{10^4}{2}\begin{bmatrix} 10^{-4} & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1/2 & -10^4/2 \\ 0 & 10^4 \end{bmatrix},$$

resulta $cond_\infty(A) = ||A||_\infty \times ||A^{-1}||_\infty = 3 \times 10^4$.
O sistema dado é mal condicionado visto que um pequeno erro relativo $||\delta_b||_\infty = 2 \times 10^{-4}/3$ no segundo membro origina um erro relativo de 200 % na solução, conforme mostrado na alínea anterior.

A.2.8

1) Considere o sistema de equações lineares

$$\begin{cases} 3\,x_1 - x_2 - x_3 & = 1 \\ x_1 - x_2 & = 0 \\ x_1 & = 5, \end{cases}$$

cuja solução é $x = (5, 5, 9)$.

(a) Escreva um sistema equivalente, de modo que o método de Jacobi seja aplicável. Justifique. Obtenha a fórmula iteradora que lhe permite calcular aproximações de x por esse método. **[1.0]**

(b) Fazendo $x^{(0)} = (1, 1, 1)$, calcule exactamente $||x - x^{(3)}||_\infty$, sendo $x^{(3)}$ a terceira iterada do método de Jacobi. **[1.0]**

(c) Diga, justificando, se o método de Jacobi converge para x, no caso de usar como aproximação inicial $x^{(0)} = (d+1, 0, d)$, sendo d o último dígito do seu número de aluno. **[1.0]**

2) Seja $h(x) = \cos\left(\frac{\pi\, x}{3}\right)$.

(a) Efectuando cálculos exactos determine o polinómio que interpola a função h nos pontos $-1, 0, 1$ e 2 . **[1.0]**

(b) Use o polinómio anterior para estimar o valor de $h(5\,\pi/24)$, e obtenha um majorante do respectivo erro de interpolação. **[1.5]**

(c) Determine a melhor aproximação de h, no sentido de mínimos quadrados, por uma função do tipo $g(x) = a + b\, x^2$, usando os pontos $-1, 0, 1$ e 2 . **[1.5]**

(d) Calcule um valor aproximado de $\displaystyle\int_0^1 \sqrt{1 + [g(x)]^2}dx$, usando 4 subintervalos de igual comprimento, e a regra de Simpson. Comece por escrever a expressão que lhe permite obter o valor pretendido. Nos cálculos utilize valores tabelados com pelo menos 4 algarismos significativos. **[1.5]**
[Caso não tenha resolvido a alínea anterior faça $g(x) = h(x)$] .

3) Pretende-se construir uma regra de quadratura $Q(f) = c_1\, f(c_2)$ para aproximar o integral $I(f) = \int_a^b f(x)dx$, onde f é uma função integrável dada. Determine as constantes c_1 e c_2 de modo que a regra tenha grau de precisão 1. Justifique. **[1.5]**

(Teste 23 Maio 2013, MEEC)

Resolução

1 (a) A matriz do sistema dado possui na sua diagonal principal uma entrada nula, o que impossibilita a aplicação do método de Jacobi nesse sistema. No entanto, um sistema equivalente ao dado é

$$\begin{cases} x_1 & = 5 \\ x_1 - x_2 & = 0 \\ 3\, x_1 - x_2 - x_3 & = 1\ . \end{cases}$$

Assim, as fórmulas iterativas para o método em causa escrevem-se,

$$\begin{cases} x_1^{(k+1)} = 5 \\ x_2^{(k+1)} = x_1^{(k)} \\ x_3^{(k+1)} = 3\, x_1^{(k)} - x_2^{(k)} - 1 \end{cases} \qquad k = 0, 1, \ldots$$

1 (b) Sendo $x^{(0)} = (1, 1, 1)$, obtém-se

$$x^{(1)} = (5, 1, 1)$$
$$x^{(2)} = (5, 5, 13)$$
$$x^{(3)} = (5, 5, 9) = x \ .$$

Logo, $||x - x^{(3)}||_\infty = 0$.

1 (c) Das fórmulas de iteração em 1 a) obtém-se imediatamente a respectiva matriz de iteração:

$$C_J = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 3 & -1 & 0 \end{bmatrix} .$$

É óbvio que $\rho(C_J) = 0$, pelo que o método converge para a solução x qualquer que seja a aproximação inicial $x^{(0)}$ escolhida (o processo dá exactamente a solução quando muito após 3 iterações).

2 (a) A partir da tabela de diferenças divididas

x_1	$h(x_i)$	$h[..]$	$h[...]$	$h[....]$
-1	$1/2$			
		$1/2$		
0	1		$-1/2$	
		$-1/2$		$1/12$
1	$1/2$		$-1/4$	
		-1		
2	$-1/2$			

obtém-se o polinómio interpolador de Newton, de grau 3,

$$\begin{aligned} p(x) &= 1/2 + 1/2\,(x+1) - 1/2\,(x+1)\,x + 1/12\,(x+1)\,x(x-1) \\ &= x^3/12 - x^2/2 - x/12 + 1 \ . \end{aligned}$$

2 (b) Seja $\bar{x} = 5\,\pi/24$. Tem-se $h(\bar{x}) \simeq p(\bar{x}) \simeq 0.754638$.
Como $h^{(4)}(x) = (\pi/3)^4 \times \cos(\pi/3\,x)$, fazendo

$$M = max_{-1\le x\le 2}\,|h^{(4)}(x)| = max_{-1\le x\le 2}\,(\pi/3)^4|\times\cos(\pi/3\,x)| = (\pi/3)^4 = \pi^4/81,$$

resulta

$$|e_{\bar{x}}| = |h(\bar{x}) - p(\bar{x})| \le \frac{M}{24}\,|\bar{x}+1|\,|\bar{x}|\,|\bar{x}-1|\,|\bar{x}-2| \simeq 0.025 \ .$$

De facto, $h(5\pi/24) = 0.77417\cdots$, confirmando a majoração obtida.

2 (c) Fazendo $\phi_0 = (1,1,1,1)^T$ e $\phi_1 = (1,0,1,4)^T$, o vector $g = a\,\phi_0 + b\,\phi_1$ é melhor aproximação de mínimos quadrados de

$$h = (h(-1), h(0), h(1), h(2))^T = (1/2, 1, 1/2, -1/2)^T,$$

se e só se (a, b) é solução do sistema de equações normais

$$\begin{bmatrix} <\phi_0,\phi_0> & <\phi_0,\phi_1> \\ <\phi_0,\phi_1> & <\phi_1,\phi_1> \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} <\phi_0, h> \\ <\phi_1, h> \end{bmatrix}, \quad i.e.,$$

$$\begin{bmatrix} 4 & 6 \\ 6 & 18 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 3/2 \\ -1 \end{bmatrix}, \quad \text{donde,}$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{36} \begin{bmatrix} 18 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 3/2 \\ -1 \end{bmatrix} = \begin{bmatrix} 11/12 \simeq 0.916667 \\ -13/36 \simeq -0.361111 \end{bmatrix}, \quad \text{ou seja,}$$

$$g(x) \simeq 0.916667 - 0.361111\, x^2 \ .$$

2 (d) Seja $f(x) = \sqrt{1 + g^2(x)}$ e $h = 1/4 = 0.25$. Atendendo a que

x_i	$f(x_i)$
0	1.35657
1/4	1.34142
1/2	1.29727
3/4	1.22847
1	1.14396

resulta para a regra de Simpson,

$$S(f) = \frac{h}{3}\{f(0) + f(1) + 4\,[f(1/4) + f(3/4)] + 2\,f(1/2)\} = 1.28122 \ .$$

Pode verificar-se que $I(f) \simeq 1.28119$, com 6 algarismos significativos.

3. A regra considerada possui grau 1 se e só se $Q(1) = I(1)$, $Q(x) = I(x)$ e $Q(x^2) \neq I(x^2)$. Ora, $c_1 = I(1) = \displaystyle\int_a^b dx = b - a$, e

$$c_1\, c_2 = I(x) \Leftrightarrow c_2 = \frac{I(x)}{c_1} = \frac{\displaystyle\int_a^b x\, dx}{b-a} = \frac{b^2 - a^2}{2\,(b-a)} = \frac{a+b}{2} \ .$$

Assim,

$$Q(f) = (b-a)\, f(\frac{a+b}{2}) \ .$$

Atendendo a que $Q(x^2) = \dfrac{(b-a)\,(a+b)^2}{4} \neq I(x^2)$, a regra é de grau 1 de exactidão.

A.2.9

[**1.**] Considere o sistema de ponto flutuante $FP(10, 5, -99, 99)$, com arredondamento simétrico.

a) Calcule $\sqrt{9.876} - \sqrt{9.875}$ nesse sistema. **[0.5]**

b) Escreva uma expressão numérica a partir da qual poderia calcular a diferença em **a)**, sem ocorrência de cancelamento subtractivo. Justifique. **[1.0]**

[**2.**] Considere a função $g(x) = (\lambda + 1)\, x - \lambda\, x^2$, onde λ é um número real não nulo.

a) Obtenha os pontos fixos da função g. **[1.0]**

b) Para cada um dos pontos fixos obtidos na alínea anterior, determine:

i. os valores de λ para os quais esses pontos fixos são atractores; **[1.0]**

ii. os valores de λ para os quais a convergência do método do ponto fixo gerado pela função g é quadrática. **[1.0]**

[**3.**] Considere a função $f(x) = (x-1)^2\, e^x$, a qual possui o zero real $z = 1$. Para o cálculo aproximado de z, pretende-se aplicar o método de Newton usual, bem como o método de Newton modificado

$$x_{n+1} = x_n - 2\,\frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, \cdots .$$

a) Mostre que z é uma raiz dupla de $f(x) = 0$. **[0.5]** .

b) O que pode dizer sobre a aplicabilidade do método da bissecção ao cálculo aproximado de $z = 1$? **[1.0]**

c) Escolhendo x_0 suficientemente próximo de z, pode garantir convergência do método de Newton? No caso afirmativo, qual é a ordem de convergência deste método? **[1.5]**

d) Pode garantir convergência do método de Newton <u>modificado</u>, no caso de escolher como aproximação inicial $x_0 = 2$? Justifique. **[1.0]**

[**4.**] Comparando a solução $x = (x_1, x_2)$ do sistema

$$\begin{cases} x_1 + 0.98\, x_2 = 4.95 \\ x_1 + x_2 = 5.0, \end{cases}$$

com a solução $\bar{x} = (\bar{x_1}, \bar{x_2})$ do sistema

$$\begin{cases} \bar{x_1} + 0.99\, \bar{x_2} = 4.95 \\ \bar{x_1} + \bar{x_2} = 5.0, \end{cases}$$

determine o erro relativo da solução, na norma $||.||_\infty$.
O que pode dizer a respeito do condicionamento do sistema? Justifique. **[1.5]**

(Exame 27 de Junho 2013, MEC, LEGM, MEAmb, LMAC)

<u>Resolução</u>

[1] (a) Sejam

$$a = \sqrt{9.876} \simeq 3.1426103\cdots, \quad b = \sqrt{9.875} \simeq 3.1424512\cdots$$
$$x = a - b = 0.0001591\cdots .$$

No sistema em causa

$$\bar{a} = fl(a) = +0.31426 \times 10^1, \qquad \bar{b} = fl(b) = +0.31425 \times 10^1$$
$$\bar{x} = fl(\bar{a} - \bar{b}) = fl(0.0001) = +0.10000 \times 10^{-3} \quad \text{(0 algarismos significativos)}.$$

[1] (b) O efeito de cancelamento subtractivo observado na alínea anterior é minorado tendo em atenção que

$$a - b = \frac{a^2 - b^2}{a + b} \ .$$

Donde,

$$\sqrt{9.876} - \sqrt{9.875} = \frac{0.001}{\sqrt{9.876} + \sqrt{9.875}} \simeq 0.0001591 \cdots \ .$$

[2] (a) Os pontos fixos da função g satisfazem a equação $g(x) = x$. Ora

$$g(x) = x \ \Leftrightarrow \lambda x + x = x + \lambda\, x^2 \Leftrightarrow x = x^2.$$

Assim, A função g tem como pontos fixos

$$\{z_1, z_2\} = \{0, 1\} \ .$$

[2] (b) i) A função $g \in C^\infty(\mathbb{R})$. Um ponto fixo z é atractor se e só se $|g'(z)| < 1$. Ora,

$$g'(x) = \lambda + 1 - 2\,\lambda\, x \ .$$

Assim,

$$\begin{array}{lll} g'(0) & = \lambda + 1 \ \Rightarrow & z_1 = 0 \text{ atractor se e só se } -2 < \lambda < 0 \\ g'(1) & = 1 - \lambda \ \Rightarrow & z_2 \text{ atractor se e só se } 0 < \lambda < 2 \ . \end{array}$$

[2] (b) ii) Escolhendo x_0 suficientemente próximo de um ponto fixo z, sabemos que se $g'(z) = 0$ e $g''(z) \neq 0$, o método de ponto fixo possuirá convergência local quadrática. Em relação ao ponto fixo $z = 0$, se fizermos $\lambda = -1$, temos

$$g'(0) = \lambda + 1 \text{ e } \ g''(0) = -2\lambda \neq 0,$$

pelo que a convergência é quadrática.
Quanto a $z = 1$, a convergência será quadrática no caso de $\lambda = 1$, já que, neste caso,

$$g'(1) = 1 - \lambda = 0 \quad \text{e} \quad g''(1) \neq 0 \ .$$

[3] (a) A função f é continuamente diferenciável quantas vezes quantas se queira.

$$\begin{array}{ll} f'(x) & = e^x \left(2\,(x-1) + (x-1)^2\right) = e^x \left(x^2 - 1\right) \\ f''(x) & = e^x \left(x^2 - 1 + 2\,x\right) \ . \end{array}$$

Assim, $f(1) = 0$, $f'(1) = 0$ e $f''(1) = 2\,e \neq 0$. Logo, $z = 1$ é zero duplo da função f.

[3] (b) A função f é contínua e não negativa. Portanto, não existe nenhum intervalo $[a, b]$, contendo o ponto $z = 1$, tal que $f(a) \times f(b) < 0$, pelo que o método da bissecção não é aplicável.

[3] (c) A função iteradora de Newton é $g(x) = x - f(x)/f'(x)$. Substituindo pela expressão da função dada, e após simplificações, obtém-se

$$\begin{aligned} g(x) &= \frac{x^2+1}{1+x} \Rightarrow g(1) = 1 \\ g'(x) &= \frac{x^2+2\,x-1}{(1+x)^2} \Rightarrow g'(1) = 1/2 \neq 0\ . \end{aligned}$$

Por conseguinte, $z = 1$ é ponto fixo atractor para g, o que significa que o método de Newton é localmente convergente, e a convergência é linear, com

$$\lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|} = |g'(1)| = \frac{1}{2},$$

uma vez escolhido x_0 suficientemente próximo de $z = 1$.

[3] (d) Neste caso, para

$$\begin{aligned} g(x) &= x - 2\,\frac{f(x)}{f'(x)} = \frac{x^2+1}{1+x} \Rightarrow g(1) = 1 \\ g'(x) &= \frac{x^2+2\,x-3}{(1+x)^2} \Rightarrow g'(1) = 0 \\ g''(x) &= \frac{8}{(1+x)^3} \Rightarrow g''(1) \neq 0\ . \end{aligned}$$

O ponto fixo $z = 1$ é superatractor, e a convergência será quadrática. De facto, efectuando, por exemplo, 4 iterações,

$$\begin{aligned} x_0 &= 2 \\ x_1 &= 1.33333333333 \\ x_2 &= 1.04761904762 \\ x_3 &= 1.00110741971 \\ x_4 &= 1.00000061285, \end{aligned}$$

evidencia-se convergência para $z = 1$, com duplicação aproximada do número de algarismos significativos de iterada para iterada, o que é típico de um método de segunda ordem de convergência.

[4] Seja $A\,x = b$ o primeiro sistema. Tem-se,

$$x = \frac{1}{0.02}\begin{bmatrix} 1 & -0.98 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 4.95 \\ 5.0 \end{bmatrix} = \begin{bmatrix} 50 & -49 \\ -50 & 50 \end{bmatrix}\begin{bmatrix} 4.95 \\ 5.0 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix}.$$

Designando por $\bar{A}\,\bar{x} = b$ o segundo sistema, resulta

$$\bar{x} = \frac{1}{0.01}\begin{bmatrix} 1 & -0.99 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 4.95 \\ 5.0 \end{bmatrix} = \begin{bmatrix} 100 & -99 \\ -100 & 100 \end{bmatrix}\begin{bmatrix} 4.95 \\ 5.0 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \end{bmatrix}.$$

Como

$$||x||_\infty = 2.5 \quad \text{e} \quad ||x - \bar{x}||_\infty = ||(2.5, 2.5)||_\infty = 2.5,$$

obtém-se,

$$||\delta_{\bar{x}}||_\infty = \frac{||x-\bar{x}||_\infty}{||x||_\infty} = 1 = 100\,\% \ .$$

O sistema é mal condicionado. Com efeito, um pequeno erro relativo numa entrada da matriz A, de grandeza $0.01/0.98 \simeq 10^{-2}$, origina um erro relativo na solução de $100\,\%$.

A.2.10

[**1.**] Considere a resolução, pelo método de Newton, do seguinte sistema de equações não lineares

$$\begin{cases} x_2^2\, e^{x_1} + x_3^2 - 2\, x_3 = 2 \\ 2\, x_2 + x_3^2 = 0 \\ x_2\, e^{2\, x_1} + x_3^2 + 6\, x_3 = -2, \end{cases}$$

tomando como aproximação inicial $x^{(0)} = (0, 1, 0)$.

a) Mostre que o sistema linear a ser resolvido para obter a primeira iterada $x^{(1)}$, é da forma

$$A\, y = b, \qquad \text{com} \quad A = \begin{bmatrix} 1 & 2 & -2 \\ 0 & 2 & 0 \\ 2 & 1 & 6 \end{bmatrix}. \qquad \textbf{[1.0]}$$

b) Determine a iterada $x^{(1)}$, efectuando cálculos exactos. **[0.5]**

c) Pode garantir a convergência do método de Jacobi para a solução do sistema $A\, y = b$,
partindo de $y^{(0)} = (27, 6, 2013)$? Justifique. **[1.5]**

[**2.**] Considere a seguinte tabela de valores de uma função $f(x)$

x_i	1	2	3	5
$f(x_i)$	0.9	0.7	0.6	0.5

a) Utilizando a fórmula de Newton com diferenças divididas, determine uma expressão para o polinómio p, de menor grau e interpolador de f, nos 3 nós mais próximos de 4. Calcule um valor aproximado de $f(4)$. **[1.5]**

b) Supondo que **[1.0]**

$$\max_{x \in R} \left| f^{(s)}(x) \right| \leq \left(\frac{\pi}{2}\right)^s, \quad s \in N,$$

apresente um majorante para o erro absoluto que se comete ao aproximar $f(2.5)$ por $p(2.5)$.

[**3.**] a) Calcule exactamente o erro de quadratura da regra de Simpson, quando aplicada

a $\int_{-1}^{1} t^4\,dt$. Qual o grau de precisão dessa regra? Justifique. **[1.5]**

b) Obtenha um valor aproximado de $I = \int_1^5 t^2 f(t)dt$, utilizando a regra de Simpson, sendo f a função tabelada em **[2.]**. Obs: use o valor $f(4) = 8/15$. **[1.5]**

[**4.**] Considere o problema de valor inicial

$$y^{'}(t) = t + sen(y(t)), \qquad y(0) = 1, \qquad t \in [0, 1] \ .$$

Utilize o método de Heun, com $h = 0.2$, para obter um valor aproximado de $y(0.2)$. Comece por escrever a fórmula de recorrência do método, aplicado ao problema em causa. **[1.5]**

(Exame 27 de Junho 2013, MEC, LEGM, MEAmb, LMAC)

Resolução

[1] (a) Dado que para $F = (f_1, f_2, f_3)$, sendo

$$\begin{array}{l} f_1(x_1, x_2, x_3) = x_2^2\, e^{x_1} + x_3^2 - 2\, x_3 - 2 \\ f_2(x_1, x_2, x_3) = 2\, x_2 + x_3^2 \\ f_3(x_1, x_2, x_3) = x_2\, e^{2\, x_1} + x_3^2 + 6\, x_3 + 2, \end{array}$$

a matriz Jacobiana de F é dada por,

$$J_F(x_1, x_2, x_3) \left[\begin{array}{ccc} x_2^2\, e^{x_1} & 2\, x_2\, e^{x_1} & 2\, x_3 - 2 \\ 0 & 2 & 2\, x_3 \\ 2\, x_2\, e^{2\, x_1} & e^{2\, x_1} & 2\, x_3 + 6 \end{array} \right].$$

Assim,

$$A = J_F(0, 1, 0) = \left[\begin{array}{ccc} 1 & 2 & -2 \\ 0 & 2 & 0 \\ 2 & 1 & 6 \end{array} \right].$$

[1] (b) O segundo membro do sistema a resolver é

$$b = -F(0, 1, 0) = (1, -2, -3)^T.$$

Aplicando o método de eliminação de Gauss ao sistema $A\, y = b$, obtém-se

$$y = (7/5, -1, -4/5)^T.$$

Por conseguinte a primeira iterada do método obtém-se resolvendo o sistema $A\, \Delta x^0 = b$, com $\Delta x^{(0)} = x^{(1)} - x^{(0)} = y$. Ou seja,

$$x^{(1)} = x^{(0)} + y = (7/5, 0, -4/5)^T.$$

[1] (c) A partir da matriz A, podemos imediatamente escrever a matriz de iteração do método de Jacobi, $C_J = -D^{-1}\,(L + U)$,

$$C_J = \left[\begin{array}{ccc} 0 & -2 & 2 \\ 0 & 0 & 0 \\ -1/3 & -1/6 & 0 \end{array} \right].$$

Visto que para as normas usuais teremos $||C_J|| > 1$, calcule-se o respectivo raio espectral.

$$Det(C_J - \lambda\, I) = \begin{vmatrix} -\lambda & -2 & 2 \\ 0 & \lambda & 0 \\ -1/3 & -1/6 & -\lambda \end{vmatrix} = -\lambda^3 - \frac{2}{3}\,\lambda\ .$$

Assim,

$$Det(C_J - \lambda\, I) = 0 \quad \text{se e só se} \quad \lambda = \pm i\,\sqrt{\frac{2}{3}} \Rightarrow \rho(C_J) = \sqrt{\frac{2}{3}} < 1,$$

pelo que o método é convergente qualquer que seja a escolha que se fizer da aproximação inicial da solução do sistema.

[2] (a) Para os valores tabelados, temos a seguinte tabela de diferenças divididas:

x_i	f_i	$f[..]$	$f[...]$
2	0.7		
		-0.1	
3	0.6		$0.1/6$
		$-0.1/2$	
5	0.5		

O polinómio interpolador é,

$$p(x) = 0.7 - 0.1(x-2) + 0.1/6\,(x-2)\,(x-3) \quad \Rightarrow p(4) = 8/15 \simeq 0.5333\cdots\ .$$

[2] (b) Aplicando a fórmula do erro de interpolação,

$$\begin{aligned} |f(x) - p(2.5)| \ &\leq \frac{1}{3!} max_{2\leq x\leq 5}\,|f^{(3)}(x)|\,|(2.5-2)\,(2.5-3)\,(2.5-5)| \\ &\leq \frac{(\pi/2)^3}{3!} \times 0.5 \times 0.5 \times 2.5 \simeq 0.403\ . \end{aligned}$$

[3] (a) Para $f(t) = t^4$, tem-se que $f^{(4)}(t) = 4!$. Por exemplo, no intervalo $[-1, 1]$, para o passo $h = 1$, o erro da regra de Simpson é,

$$E_S(f) = -\frac{2}{180} \times 4! = -\frac{4}{15} \neq 0\ .$$

Por conseguinte a regra não é exacta para polinómios de grau 4 mas, por construção, é exacta para qualquer polinómio de grau ≤ 3. Logo, a regra é de grau 3.

[3] (b) Seja $F(t) = t^2\, f(t)$. Para a regra de Simpson com passo $h = 1$, serão usados os valores da tabela

t_i	1	2	3	4	5
$F(t_i)$	0.9	2.8	5.4	128/15	12.5

A aproximação de I pretendida é

$$S(F) = \frac{1}{3}\,[F(1) + F(5) + 4\,(F(2) + F(4)) + 2\,F(3)] \simeq 23.1778\ .$$

[4] Sendo $f(t,y) = t + \sin(y)$, passo h, e nós $t_i = i\,h$, $i = 0, 1, \ldots$, a fórmula de recorrência do método é

$$y_{i+1} = y_i + \frac{h}{2}\,[f(t_i, y_i) + f(\,t_i + h, y_i + h\,f(t_i, y_i)\,)]\,.$$

Para $t_0 = 0$, $y_0 = 1$ e $h = 0.2$, obtém-se

$$y(0.2) \simeq y_1 = 1 + \frac{0.2}{2}\,[\sin(1) + 0.2 + \sin(1 + 0.2 \times \sin(1))] \simeq 1.19616\ .$$

A.2.11

I

Considere a função iteradora,

$$g(x) = k\,x(1-x), \quad \text{com} \quad k > 0\ .$$

1. Determine os pontos fixos de g (em função de k). **[1.0]**

2. No caso de $1 < k < 2$, diga se cada um dos pontos fixos é atractor ou repulsor, **[1.0]** justificando a resposta.

3. Seja $k = 1.5$. Considere a sucessão $\{x_n\}$, definida por :

$$x_0 = 0.5, \quad x_{n+1} = g(x_n), \quad n = 0, 1, 2, \ldots\ .$$

Diga se esta sucessão converge, apresentando justificação teórica. Em caso afirmativo, **[1.0]** diga qual o seu limite.

4. Para $k = 1.5$, pretende-se aproximar os pontos fixos de g, usando o método de Newton. Mostre que, neste caso, se obtém a função iteradora

$$h(x) = \frac{1.5\,x^2}{3\,x - 0.5}\ .$$

[1.0]

5. Partindo de $x_0 = 0.5$, efectue as duas primeiras iterações do método referido na alínea anterior. Como compara este método com o da alínea 3, quanto à rapidez de **[1.0]** convergência? (Baseie a sua resposta no conhecimento teórico sobre esses métodos).

II

Considere um sistema linear $\mathbf{A}\mathbf{x} = \mathbf{b}$, onde

$$\mathbf{A} = \begin{bmatrix} 3 & a & 0 \\ a & 3 & a \\ 0 & a & 3 \end{bmatrix}.$$

[1.0]

1. (i) Diga (justificando) para que valores de a o sistema é mal condicionado.
   Obs: tenha em conta que a inversa de A, quando existe, é dada por

$$\mathbf{A^{-1}} = \frac{1}{27-6a^2}\begin{bmatrix} 9-a^2 & -3a & a^2 \\ -3a & 9 & -3a \\ a^2 & -3a & 9-a^2 \end{bmatrix}.$$

   (ii) Diga o que entende por um sistema mal condicionado. **[0.5]**

2. Indique um intervalo $J = [\alpha, \beta]$, de modo que o método iterativo de Jacobi, aplicado a um sistema $A\,x = b$, seja convergente se e só se o parâmetro $a \in J$. Justifique. **[1.5]**

3. Considere $a = -1$. Prove que as iteradas do método de Jacobi satisfazem **[1.0]**

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq 2\,\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty, \qquad k = 0, 1, 2, \ldots$$

4. Seja $b = (2, 1, 2)$. Tomando como aproximação inicial o vector $x^{(0)} = (1, 1, 1/2)$, efectue duas iterações do método de Jacobi. Usando a estimativa da alínea anterior, obtenha um majorante de $\|\mathbf{x}^{(2)} - \mathbf{x}\|_\infty$. **[1.0]**

III

Considere uma função de variável real f, tal que

$$f(1) = 1, \quad f(x) = f(x-2) + (x-1)^2, \quad x > 0\ .$$

1. Determine o polinómio que interpola f em $x = 1$, $x = 3$ e $x = 5$. **[1.5]**

2. Mostre que

$$f[x, x+2, x+4, x+6] = \frac{1}{6}, \quad \forall x \geq 1\ .$$

   Com base nesta igualdade, e admitindo que $f \in C^3([1, \infty[)$, mostre que f é um polinómio e determine o seu grau. **[1.5]**

3. Determine um valor aproximado de $\int_1^9 (x-3) f(x)dx$, usando a regra de Simpson composta. **[1.5]**

4. Tendo em conta o que foi provado na alínea 2, determine o erro absoluto da aproximação obtida na alínea 3. (Se não resolveu a alínea 2, assuma que f é um polinómio de grau 3, $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$) . **[1.5]**

IV

1. Considere o problema de valor inicial

$$\begin{cases} y'(x) = -2\sin\left((x+1)\,y(x)\right), \quad 0 \leq x \leq 1 \\ y(0) = 1\ . \end{cases}$$

   a) Aplique o método de Euler, com passo $h = 0.1$, e calcule uma aproximação para $y(0.2)$. **[1.0]**

   b) Obtenha um majorante para o erro absoluto do valor obtido na alínea anterior. **[1.5]**

2. Utilizando um determinado método numérico, designado por MN, foram obtidas soluções aproximadas de um problema de valor inicial para uma equação diferencial ordinária de primeira ordem. Na seguinte tabela estão apresentadas as aproximações obtidas, usando diferentes valores do passo h, bem como a solução exacta:

h	$y(1)$
0.5	0.8234
0.25	0.8162
sol. exacta	0.8090

[1.0] a) Diga, justificando, qual dos seguintes métodos pode ter sido usado para obter estas aproximações: i) Euler explícito; ii) Taylor de segunda ordem.

[0.5] b) Que valor espera obter, se usar o método MN com passo $h = 0.125$?

(Exame de 15 de Julho de 2013)

Resolução

I

1. Pretende-se resolver a equação

$$g(z) = z \Leftrightarrow kz(1-z) = z \Leftrightarrow z = 0 \vee k(1-z) = 1 \Leftrightarrow z = 0 \vee k(1-z) = 1$$
$$\Leftrightarrow z = 0 \vee z = 1 - 1/k \ .$$

Existem, portanto, dois pontos fixos $z_1 = 0$ e $z_2 = 1 - 1/k$.

2. Temos $g'(z) = k - 2\,k\,z$. Para $z_1 = 0$, resulta $g'(z_1) = k$. Logo, $|g'(z_1)| > 1$, pelo que z_1 é um ponto repulsor.
Para $z_2 = 1-1/k$, temos $g'(z_2) = 2-k$. Visto que que $1 < k < 2$, resulta $0 < g'(z_2) < 1$, pelo que z_2 é ponto fixo atractor.

3. Se a sucessão convergir, será para o ponto fixo atractor de g, z_2, que neste caso é $z_2 = 1 - 1/k = 1/3$.

Veriquemos as condições suficientes de convergência do teorema do ponto fixo no intervalo $I = [1/3, 1/2]$.

(i) $g(I) \subset I$. Para mostrar esta condição, comecemos por verificar se g é monótona em I . Temos $g'(x) = 1.5 - 3\,x \geq 0, \forall x \in I$. Consequentemente, g é monótona (crescente) em I . Além disso, $g(1/3) = 1/3 \in I$ e $g(1/2) = 3/8 \in I$. Logo, a condição mencionada é satisfeita.

(ii) $g \in C^1(I)$ e $\max_{x\in I} |g'(x)| < 1$. A primeira parte desta condição é evidente, visto que $g'(x) = 1.5 - 3\,x$. Quanto à segunda parte, sendo $g'(x)$ não negativa e decrescente em I, temos

$$\max_{x\in I} |g'(x)| = g'(1/3) = 0.5 < 1 \ .$$

Finalmente, pelo teorema do ponto fixo, a sucessão converge.

4. Para determinarmos os pontos fixos de g devemos resolver a equação $f(z) = g(z) - z = 0$. Neste caso,

$$f(x) = 1.5\,x\,(1-x) - x = 0.5\,x - 1.5\,x^2 \ .$$

Para obtermos aproximações dos zeros de f pelo método de Newton devemos considerar a função iteradora

$$h(x) = x - f(x)/f'(x) = x - \frac{0.5\,x - 1.5\,x^2}{0.5 - 3\,x} = \frac{1.5\,x^2}{3\,x - 0.5} \ .$$

5. Quanto à rapidez de convergência, em primeiro lugar, deveremos mostrar que o método de Newton converge quando fazemos $x_0 = 0.5$. Por exemplo, no intervalo $I = [0.25, 0.5]$, tem-se:
f é contínua, $f(0.25) > 0$ e $f(0.5) < 0$;
$f'(x) = 0.5 - 3x$ é negativa em I;
$f''(x) = -3 < 0$ em I;
$f(0.5) \times f''(x) \geq 0$, $\forall x \in I$.
As quatro condições anteriores garantem que no caso considerado o método de Newton converge, e a sua convergência é quadrática para $z = 1/3 = 0.3333\cdots$. O método do ponto fixo, considerado na alínea 2, possui convergência apenas linear, visto que $g'(z_2) = 0.5 \neq 0$. Logo, o método de Newton é mais rápido.

Para $x_0 = 0.5$, obtém-se

$$\begin{aligned} x_1 &= h(x_0) = 0.375 \\ x_2 &= h(x_1) = 0.3375 \ . \end{aligned}$$

II

1. (i) Antes de mais, precisamos de calcular $\|A\|$ e $\|A^{-1}\|$. Escolhendo a norma por linha (por exemplo), temos

$$||A||_\infty = \max(|a| + 3, 2|a| + 3) = 2|a| + 3$$

$$||A^{-1}||_\infty = \tfrac{1}{|27-6a^2|} \max(|9 - a^2| + |3a| + |a^2|, |6a| + |9|) \ .$$

Assim, verifica-se que $\|A\|_\infty \to \infty$, sse $|a| \to \infty$.
Por outro lado, $\|A^{-1}\|_\infty \to \infty$, sse $|27 - 6\,a^2| \to 0$, ou seja, sse $a \to \pm\sqrt{9/2}$. Basta, portanto, analisar estes dois valores de a .

No caso de $|a| \to \infty$, temos $\|A\|_\infty \to \infty$ e

$$\begin{aligned} &\lim_{|a|\to\infty} \|A^{-1}\|_\infty = \lim_{|a|\to\infty} \frac{1}{|27-6a^2|} \max(|9 - a^2| + |3a| + |a^2|, |6a| + |9|) \\ &= \max(\lim_{|a|\to\infty} \frac{|9 - a^2| + |3a| + |a^2|}{|27 - 6a^2|}, \\ &= \lim_{|a|\to\infty} \frac{9 + |6a|}{|27 - 6a^2|}) = \max(\frac{1}{3}, 0) = \frac{1}{3} \ . \end{aligned}$$

Por conseguinte, $cond(A)$ tende para infinito e o sistema é mal condicionado.

No caso de $|a| \to \sqrt{\frac{9}{2}}$, temos

$$\|A\|_\infty \to 3 + 2\sqrt{\frac{9}{2}},$$

$$\|A^{-1}\|_\infty \to \infty,$$

logo, $cond(A)$ tende para infinito e o sistema também é mal condicionado. Para outros valores de a o sistema é bem condicionado.

(ii) Um sistema mal condicionado é aquele em que pequenos erros relativos na matriz ou no segundo membro podem provocar grandes erros relativos na solução. O condicionamento de um sistema pode ser verificado através do número de condição da sua matriz,

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

No caso do sistema dado, ele será mal condicionado para um certo valor de a, se $cond(A)$ tender para infinito, quando a tende para esse valor.

2. É condição necessária e suficiente de convergência do método de Jacobi que $\rho(C) < 1$, onde $\rho(C)$ representa o raio espectral da matriz de iteração do método. Temos

$$\mathbf{C} = \begin{bmatrix} 0 & -a/3 & 0 \\ -a/3 & 0 & -a/3 \\ 0 & -a/3 & 0 \end{bmatrix} \ .$$

A equação característica de C é

$$Det(C - \lambda I) = -\lambda^3 + 2\,\lambda\, a^2/9 = 0,$$

pelo que os respectivos valores próprios são

$$\lambda_1 = 0, \qquad \lambda_{2,3} = \pm a\frac{\sqrt{2}}{3} \ .$$

Logo, $\rho(C) = |a|\frac{\sqrt{2}}{3}$. A condição a satisfazer é

$$\rho(C) < 1 \Leftrightarrow |a|\frac{\sqrt{2}}{3} < 1 \Leftrightarrow |a| < \frac{3}{\sqrt{2}},$$

pelo que o intervalo pedido é $]-\frac{3}{\sqrt{2}}, \frac{3}{\sqrt{2}}[$.

3. Para obter a estimativa do erro, em primeiro lugar, temos que calcular $\|C\|_\infty$,

$$\|C\|_\infty = \max(1/3, 2/3, 1/3) = 2/3 \ .$$

Assim,

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \frac{\|C\|_\infty}{1 - \|C\|_\infty}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty = 2\,\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty \ .$$

4. Primeira iteração do método de Jacobi:

$$\begin{aligned} x_1^{(1)} &= \frac{2 + x_2^{(0)}}{3} = 1 \\ x_2^{(1)} &= \frac{1 + x_1^{(0)} + x_3^{(0)}}{3} = 5/6 \\ x_3^{(1)} &= \frac{2 + x_2^{(0)}}{3} = 1 \ . \end{aligned}$$

Segunda iteração:

$$\begin{aligned} x_1^{(2)} &= \frac{2 + x_2^{(1)}}{3} = 17/18 \\ x_2^{(2)} &= \frac{1 + x_1^{(1)} + x_3^{(1)}}{3} = 1 \\ x_3^{(2)} &= \frac{2 + x_2^{(1)}}{3} = 17/18 \ . \end{aligned}$$

Estimativa de erro:

$$\|\mathbf{x} - \mathbf{x}^{(2)}\|_\infty \leq 2 \, \|\mathbf{x^{(2)}} - \mathbf{x}^{(1)}\|_\infty = 2 \times 1/6 = 1/3 \ .$$

III

1. Calculemos os valores de f nos pontos a considerar,

$$\begin{aligned} f(3) &= f(1) + 2^2 = 1 + 4 = 5 \\ f(5) &= f(3) + 4^2 = 5 + 16 = 21, \end{aligned}$$

e as diferenças divididas:

$$\begin{aligned} f[1,3] &= (5-1)/2 = 2 \\ f[3,5] &= (21-5)/2 = 8 \\ f[1,3,5] &= (8-2)/4 = 3/2 \ . \end{aligned}$$

Pela fórmula interpoladora de Newton:

$$P_2(x) = 1 + 2(x-1) + 3/2(x-1)(x-3) \ .$$

2. Seja $x > 0$ um número real arbitrário. Atendendo à definição da função f, temos

$$\begin{aligned} &f[x, x+2] = (f(x+2) - f(x))/2 = (x+1)^2/2 \\ &f[x+2, x+4] = (f(x+4) - f(x+2))/2 = (x+3)^2/2 \\ &f[x+4, x+6] = (f(x+6) - f(x+4))/2 = (x+5)^2/2 \\ &f[x, x+2, x+4] = 1/4 \left((x+3)^2/2 - (x+1)^2/2\right) = (2x+4)/4 \\ &f[x+2, x+4, x+6] = 1/4 \left((x+5)^2/2 - (x+3)^2/2\right) = (2x+8)/4 \\ &f[x, x+2, x+4, x+6] = 1/6 \left((2x+8)/4 - (2x+4)/4\right) = 1/6 \ . \end{aligned}$$

Fica assim provada a igualdade. Recordemos agora que se $f[x, x+2, x+4, x+6] = const$ (não depende de x) e se $f \in C^3([1, \infty[)$, então a terceira derivada de f também é

constante (igual a $1/6 \times 3! = 1$) . Daqui resulta que f é um polinómio de terceiro grau em $[1, \infty[$.

3. Para usar a regra de Simpson composta, uma vez que a função f só é conhecida nos pontos $x = 1, 3, 5, ...$, temos de considerar $h = 2$. Assim, os nós de integração são: $x_0 = 1, x_1 = 3, x_2 = 5, x_3 = 7, x_4 = 9$. Uma vez que já conhecemos os valores de $f(1), f(3), f(5)$ (alínea 1), vamos calcular $f(7)$ e $f(9)$.

$$\begin{aligned} f(7) &= f(5) + 6^2 = 21 + 36 = 57 \\ f(9) &= f(7) + 8^2 = 57 + 64 = 121 \ . \end{aligned}$$

A função integranda é $g(x) = (x-3)f(x)$. Para esta função temos

$$\begin{aligned} g(1) &= f(1)(1-3) = -2 \\ g(3) &= f(3)(3-3) = 0 \\ g(5) &= f(5)(5-3) = 42 \\ g(7) &= f(7)(7-3) = 228 \\ g(9) &= f(9)(9-3) = 726 \ . \end{aligned}$$

Aplicando a fórmula da regra de Simpson composta, obtém-se

$$S_4(g) = h/3\,(g(1) + 4\,g(3) + 2\,g(5) + 4\,g(7) + g(9)) = 3440/3 \approx 1146,67 \ .$$

4. O erro de truncatura da regra de Simpson é dado por

$$E_S^4(g) = -\frac{h^4(b-a)}{180}\, g^{(4)}(\xi), \qquad \xi \in [1, 9] \ .$$

Avaliemos a quarta derivada de g. Em primeiro lugar, sabemos que f é um polinómio de grau 3, logo g é um polinómio de quarto grau. Como vimos na alínea 2, $f[x, x+2, x+4, x+6] = 1/6$. Donde, $f(x) = x^3/6 + ...$ (onde as reticências representam os termos de graus inferiores). Finalmente, temos $g(x) = x^4/6 + ...$. Daqui se conclui que $g^{(4)}(\xi) = 4!/6 = 4$ (qualquer que seja ξ) .
Substituindo na fórmula do erro, resulta

$$E_S^4(g) = -\frac{4h^4(b-a)}{180} = -\frac{4 \times 2^4 \times 8}{180} = -128/45 \approx 2.844 \ .$$

O erro absoluto tem o valor $128/45$.

IV

1 (a). Aplicando a fórmula do método de Euler,

$$y_{i+1} = y_i + hf(x_i, y_i) = y_i - 2h\sin((x_i + 1)y_i)) \ .$$

Uma vez que $h = 0.1$, precisamos de efectuar dois passos. Temos $x_0 = 0, x_1 = 0.1, x_2 = 0.2$.
Primeiro passo,

$$y_1 = y_0 - 2\,h\sin(y_0) = 1 - 0.2\sin(1) = 0.831706 \ .$$

Segundo passo,

$$y_2 = y_1 - 2\,h\,\sin(1.1\,y_1) = 0.673208\ .$$

1 (b). Fórmula do erro do método de Euler:

$$|y(x_2) - y_2| \leq \frac{hY_2}{2}\frac{e^{x_2K}-1}{K},\ \text{onde } K = \max_{x\in[0,x_2]}\left|\frac{\partial f}{\partial y}\right| \text{ e } Y_2 = \max_{x\in[0,x_2]}|y''(x)|\ .$$

Como

$$\frac{\partial f}{\partial y} = 2(x+1)\cos((x+1)y),$$

logo

$$K = \max_{x\in[0,x_2]}|2(x+1)\cos((x+1)y)| \leq 2\times 1.2 = 2.4\ .$$

Por outro lado,

$$y''(x) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y'(x) = -2\cos((x+1)y)y - 2(x+1)\cos((x+1)y)y'\ .$$

Por conseguinte,

$$|y''(x)| \leq 2|y(x)| + 2(x+1)|y'(x)|\ .$$

Atendendo a que que y é decrescente (pois $y'(0)$ é negativo), logo $y(x) \leq 1$, donde $|y'(x)| \leq 2$ (de acordo com a equação diferencial). Finalmente, obtém-se

$$Y_2 \leq 2 + 4.8 = 6.8.$$

Assim, substituindo K e Y_2 na fórmula do erro, resulta

$$|y(x_2) - y_2| \leq h\,3.4\frac{e^{0.48}-1}{2.4} = 0.087\ .$$

2 (a). Os erros cometidos, em cada caso, são:

$$\begin{array}{ll} h = 0.5, & e_h = 0.8234 - 0.8090 = 0.0144 \\ h = 0.25, & e_h = 0.8162 - 0.8090 = 0.0072\ . \end{array}$$

Comparando os valores anteriores, verifica-se que para $h = 0.25$ o erro ficou reduzido a metade. Conclui-se assim que se trata de um método de primeira ordem.

2 (b). Uma vez que se trata de um método de primeira ordem, espera-se que, ao diminuir o passo para metade, o erro volte a reduzir-se na mesma proporção. Assim, para $h = 0.125$, deveremos ter $e_h \approx 0.0072/2 = 0.0036$. Deste modo, o valor esperado da solução é $0.8090 + 0.0036 = 0.8126$.

A.2.12

1) Considere um sistema de ponto flutuante e arredondamento simétrico, de base 10 e 4 dígitos na mantissa.

(a) Sendo k o seu número de aluno, que valor obtém se calcular **[1.0]**

$$v = \pi \, \frac{10^{-6}}{k^2}$$

nesse sistema? Indique todos os passos e cálculos do algoritmo que utilizar.

[1.0] **(b)** Diga, justificando, se a função

$$\phi(x) = k \, \frac{\sin(x)}{x},$$

para $x > 0$, é bem condicionada para valores de x próximos de zero. (A constante k designa o seu número de aluno).

2) Sabe-se que a equação $x^3 - 6\,x^2 + 9\,x - 5 = 0$ possui uma raiz z no intervalo $I = [4, 5]$. Considere as funções iteradoras

$$h_1(x) = \frac{2\,x^3 - 6x^2 + 5}{3\,x^2 - 12\,x + 9} \quad \text{e} \quad h_2(x) = -x^3 + 6x^2 - 8\,x + 5 \ .$$

[1.0] **(a)** Verifique que a função h_1 corresponde à função iteradora do método de Newton.

[1.5] **(b)** Prove que se pode assegurar a convergência do método de Newton com qualquer iterada inicial $x_0 \in I$. Indique <u>dois</u> valores possíveis para uma aproximação inicial x_0 da raiz z, para os quais se possa garantir convergência <u>monótona</u> do método. Justifique.

[1.5] **(c)** Aproxime z com erro inferior a 10^{-8}, usando o método de Newton. Justifique convenientemente usando uma majoração de erro que considere apropriada.

[1.5] **(d)** Partindo de $x_0 = 4.1$, calcule as duas primeiras iteradas do método gerado por h_2. Pode garantir que a respectiva sucessão $(x_k)_{k\geq 0}$ converge para z? Justifique teoricamente.

3) A matriz A (tridiagonal e simétrica) de um sistema (3×3), $A\,x = b$, é definida por $a_{i,i} = 3$ e $a_{i,j} = -1$, se $|i - j| = 1$. São nulas as restantes entradas da matriz. O segundo membro do sistema é dado por $b_i = \sum_{j=1}^{j=3} a_{i,j}$, para i desde 1 a 3 .

[1.5] **(a)** Efectuando cálculos <u>exactos</u>, obtenha as <u>duas</u> primeiras iteradas do método de Gauss-Seidel aplicado ao sistema. Parta do ponto $x^{(0)} = (k, 0, 0)$, onde k é o seu número de aluno.

[1.0] **(b)** Diga, justificando, se o referido método converge para a solução $(1, 1, 1)$ do sistema dado, caso escolha um ponto inicial <u>qualquer</u> $x^{(0)} \in \mathbb{R}^3$.

(Teste 7 Nov. 2013)

<u>Resolução</u>

1(a) Seja, por exemplo, $k = 75200$.

$$\begin{array}{ll} v_1 = \pi & \to \bar{v}_1 = fl(\pi) = +0.3142 \times 10^1 \\ v_2 = v_1 \times 10^{-6} & \to \bar{v}_2 = fl(\bar{v}_1 \times 10^{-6}) = +0.3142 \times 10^{-5} \\ v_3 = k & \to \bar{v}_3 = fl(k) = +0.7520 \times 10^5 \\ v_4 = v_3 \times v_3 & \to \bar{v}_4 = fl(0.565504 \times 10^{10}) = +0.5655 \times 10^{10} \\ v = \dfrac{v_2}{v_4} & \to \bar{v} = fl\left(\dfrac{\bar{v}_2}{\bar{v}_3}\right) = fl(0.555615\cdots \times 10^{-15}) = +0.5556 \times 10^{-15} \ . \end{array}$$

1(b) O número de condição de ϕ:

$$cond_\phi(x) = \left|\frac{x\,\phi'(x)}{\phi(x)}\right| = \left|\frac{x\,\cos(x) - \sin(x)}{\sin(x)}\right| = \left|\frac{x\,\cos(x)}{\sin(x)} - 1\right| .$$

Atendendo a que

$$\lim_{x\to 0} \frac{\sin(x)}{x} = 1 \, ,$$

temos

$$\lim_{x\to 0} cond_\phi(x) = \lim_{x\to 0} \left|\frac{\cos(x)}{\sin(x)/x} - 1\right| = 0 \ .$$

A função em causa é bem condicionada para $x \simeq 0$.

2(a) Para $f(x) = x^3 - 6\,x^2 + 9\,x - 5$, a função iteradora de Newton escreve-se,

$$\begin{aligned} g(x) \ &= x - f(x)/f'(x) \\ &= x - \frac{x^3 - 6\,x^2 + 9\,x - 5}{3\,x^2 - 12\,x + 9} = \frac{2\,x^3 - 6\,x^2 + 5}{3\,x^2 - 12\,x + 9} = h_1(x) \ . \end{aligned}$$

2(b) Como $f \in C^2(I)$ e

$$\begin{array}{l} f(4) \times f(5) = -15 < 0 \\ f'(x) = 3\,x^2 - 12\,x + 9 \\ f''(x) = 6\,x - 12 > 0, \quad \forall x \in I, \end{array}$$

conclui-se que f' é função estritamente crescente em I. Dado que $f'(4) = 9 > 0$, resulta que no intervalo $f'(x) > 0$. Assim, o zero z é simples e único nesse intervalo. Finalmente, verifica-se a condição que garante que os zeros das tangentes ao gráfico de f em $(4, f(4))$ e $(5, f(5))$ estão contidos no interior de I, ou seja: $|f(4)|/|f'(4)| < 5 - 4 = 1$ e $|f(5)|/|f'(5)| < 5 - 4 = 1$.

Para se obter convergência monótona, deve-se escolher x_0 de modo a verificar-se: $f(x_0)f''(x) \geq 0, \forall x \in I$. Por exemplo, para $x_0 = 5$ ou $x_0 = 4.5$,

$$f(5) = 15 > 0 \quad \text{e} \quad f(4.5) = 5.125 > 0 \quad \text{possuem o mesmo sinal de } f''.$$

Assim, o método de Newton converge (quadraticamente) para z e a convergência é monótona (com todas as iteradas à direita de z).

2(c) Com $x_0 = 4$, obtém-se

$$x_1 = 4.11111111$$

Dado que $f(x_0) < 0$ e $f(x_1) > 0$, conclui-se que $z \in [x_0, x_1]$, com

$$|z - x_k| \leq x_1 - x_0 = 0.11111111, \quad k = 0, 1 \ .$$

Além disso, todas as iteradas do método de Newton ficam no intervalo $[x_0, x_1]$ uma vez que a sucessão de iteradas é monótona a partir de x_1. Com base nestas considerações e nas propriedades de f referidas na alínea anterior, tem-se

$$\mathbb{K} := \frac{\max_{x\in[x_0,x_1]} |f''(x)|}{2\min_{x\in[x_0,x_1]} |f'(x)|} = \frac{|f''(x_1)|}{2|f'(x_0)|} = 0.703703704 \ .$$

Assim, o erro da iterada k do método, com início em $x_0 = 4$, pode ser majorado por

$$|z - x_k| \leq \frac{(\mathbb{K}|z - x_0|)^{2^k}}{\mathbb{K}} \leq \frac{(0.078189300)^{2^k}}{0.703703704}, \quad k = 0, 1, 2, ... \ .$$

Por conseguinte, a iterada x_3 satisfaz $|z - x_3| < 10^{-8}$. Calculemos então x_2 e x_3:

$$x_2 = 4.10383598, \quad x_3 = 4.10380340 \ .$$

2(d) Para $x_0 = 4.1$, as duas primeiras iteradas do método gerado por h_2 são:

$$\begin{aligned} x_1 &= h_2(4.1) \simeq 4.139 \\ x_2 &= h_2(x_1) \simeq 3.76939 \ . \end{aligned}$$

Como $h_2(z) = -z^3 + 6\,z^2 - 8\,z + 5 = -z^3 + 6z^2 - 9\,z + 5 + z$, resulta que $h_2(z) = z$, pois $f(z) = 0$. Assim, z é ponto fixo de h_2 . No entanto, a sucessão $x_{k+1} = h_2(x_k)$, para $k = 0, 1, \ldots$ não pode convergir para z . Com efeito,

$$h_2'(x) = -3\,x^2 + 12\,x - 8 \ .$$

Pode concluir facilmente que h_2' é estritamente decrescente em I e $h_2'(-4) = -8$. Por conseguinte, $|h_2'(x)| > 1$, $\forall x \in I$. Em particular, $|h'(z)| > 1$, pelo que o ponto fixo z é repulsor para esta função iteradora.

3(a) O sistema a resolver é

$$\begin{cases} 3\,x_1 - x_2 & = 2 \\ -x_1 + 3\,x_2 - x_3 & = 1 \\ -x_2 + 3\,x_3 = 2 \end{cases}$$

As respectivas fórmulas computacionais escrevem-se,

$$\begin{cases} x_1^{(k+1)} & = \dfrac{2 + x_2^{(k)}}{3} \\ x_2^{(k+1)} & = \dfrac{1 + x_1^{(k+1)} + x_3^{(k)}}{3} \\ x_3^{(k+1)} & = \dfrac{2 + x_2^{(k+1)}}{3}. \end{cases} \qquad k = 0, 1, \ldots$$

Para $x^{(0)} = (k, 0, 0)$, resulta

$$\begin{cases} x_1^{(1)} &= \dfrac{2}{3} \\ x_2^{(1)} &= \dfrac{1+\frac{2}{3}}{3} = \dfrac{5}{9} \\ x_3^{(1)} &= \dfrac{2+\frac{5}{9}}{3} = \dfrac{23}{27} \end{cases}$$

e

$$\begin{cases} x_1^{(2)} &= \dfrac{2+\frac{5}{9}}{3} = \dfrac{23}{27} \\ x_2^{(2)} &= \dfrac{1+2\frac{23}{27}}{3} = \dfrac{73}{81} \\ x_3^{(2)} &= \dfrac{2+\frac{73}{81}}{3} = \dfrac{235}{243} \, . \end{cases}$$

3(b) A matriz do sistema anterior é estritamente diagonal dominante (por linhas ou colunas) pelo que o método converge para a solução $x = (1, 1, 1)$, qualquer que seja a aproximação inicial escolhida.

A.2.13

Exame de 29 de Janeiro de 2014 (Duração: 1h30m) – Parte 1

I

Seja a equação $P(x) - e^x = 0$, onde $P(x) = 2 - x^2$.

Considere dois métodos iterativos para aproximação das raízes da equação em causa, definidos pelas fórmulas,

$$x_{n+1} = x_n - \frac{P(x_n) - e^{x_n}}{P'(x_n) - e^{x_n}} \qquad (1) \qquad\qquad x_{n+1} = x_n + \frac{P(x_n) - e^{x_n}}{2} \, . \qquad (2)$$

(a) Justifique que, no caso de $x_0 \in [0.5, 0.6]$, o método correspondente à fórmula (1) gera uma sucessão que converge para a raiz positiva da equação considerada. **[1.5]**

(b) Utilizando o método (2), com $x_0 = 0.5$, calcule x_2 e obtenha uma estimativa do erro absoluto da aproximação calculada. Mostre ainda que o método converge. **[1.5]**

(c) Com base na noção de ordem de convergência, diga qual das sucessões (1) ou (2) converge mais rapidamente. **[1.0]**

(d) É possível usar o método (2) para aproximar a raiz negativa da equação considerada? Justifique (sem fazer iterações). **[1.0]**

II

1) Considere o sistema linear $A\,x = b$, onde $x = (u, v, w)$,

$$A = \begin{bmatrix} -1 & a & 0 \\ 1/2 & -1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \quad \text{e} \qquad b = (-1, \epsilon, c), \quad \text{com} \quad \epsilon, a, c \in \mathbf{R}\ .$$

(a) Mostre que, ao resolver o sistema pelo método de Gauss-Seidel, se obtêm as seguintes fórmulas,

$$\begin{aligned} u^{(k+1)} &= a\,v^{(k)} + 1 \\ v^{(k+1)} &= -\epsilon + \frac{a\,v^{(k)} + 1}{2} \\ w^{(k+1)} &= \frac{c - a\,v^{(k)} - 1}{3}, \end{aligned}$$

[1.0] onde $(u^{(k)}, v^{(k)}, w^{(k)})$ representa a k-ésima iterada do método referido.

(b) Justifique que o método de Gauss-Seidel converge para a solução do sistema linear considerado, qualquer que seja a aproximação inicial $(u^{(0)}, v^{(0)}, w^{(0)})$, se e só se $|a| < 2$.

[1.0]

(c) Fazendo $a = 0$, mostre que se aplicar o método de Jacobi se obtém a solução exacta do sistema, no máximo ao fim de 3 iterações, qualquer que seja a aproximação inicial

[1.5] que considere.

[1.5] **2)** Seja $a \in \mathbb{R}$. Considere as matrizes,

$$A = \begin{bmatrix} -1 & a & 0 \\ 0.5 & -1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \quad e \quad A^{-1} = \begin{bmatrix} -\dfrac{3}{3 - 1.5\,a} & -\dfrac{3\,a}{3 - 1.5\,a} & 0 \\ -\dfrac{1.5}{3 - 1.5\,a} & -\dfrac{3}{3 - 1.5\,a} & 0 \\ \dfrac{1}{3 - 1.5\,a} & \dfrac{a}{3 - 1.5\,a} & \dfrac{1 - 0.5\,a}{3 - 1.5\,a} \end{bmatrix}.$$

Para a norma matricial induzida $||\,.\,||_\infty$, discuta o condicionamento da matriz A, tendo em conta os seguintes casos: (i) quando $a \simeq 2$; (ii) quando $|a|$ toma valores muito elevados e (iii) quando $a \simeq 0$.

Resolução

I

(a) A fórmula (1) corresponde à aplicação do método de Newton à função $f(x) = P(x) - e^x$. De facto, quando se aplica este método à resolução da equação $f(x) = 0$, obtém-se

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{P(x_n) - e^{x_n}}{P'(x_n) - e^{x_n}}\ .$$

Verifiquemos que no intervalo $[0.5, 0.6]$ são satisfeitas as condições suficientes de convergência do método de Newton.

(i) $f(0.5) > 0 \quad \text{e} \quad f(0.6) < 0$.
(ii)

$$f'(0.5) = -2.65 \quad \text{e} \quad f'(0.6) = -3.022,$$

(notar que f'' é negativa em $\mathbb{R}$; logo, f' é decrescente e, por conseguinte, não se anula em $[0.5, 0.6]$) .
(iii) $f''(x) < 0$ em $[0.5, 0.6]$.
(iv) $\dfrac{|f(0.5)|}{|f'(0.5)|} = 0.038 < 0.1 \quad \text{e} \quad \dfrac{|f(0.6)|}{|f'(0.6)|} = 0.06 < 0.1$.

Assim, o método (1) converge para a raiz positiva da equação, qualquer que seja $x_0 \in [0.5, 0.6]$.

(b) Para $x_0 = 0.5$, tem-se

$$x_1 = g(x_0) = x_0 + \frac{P(x_0) - e^{x_0}}{2} \simeq 0.550639$$

$$x_2 = g(x_1) = x_1 + \frac{P(x_1) - e^{x_1}}{2} \simeq 0.531857 \ .$$

Note-se,

$$g'(x) = 1 - x - \frac{e^x}{2} \quad \text{e} \quad g''(x) = -1 - \frac{e^x}{2},$$

são funções contínuas em $\mathbb{R}$ (e, em particular, no intervalo considerado). Como g'' é negativa, a função g' é decrescente.

Atendendo a que $g'(0.5) \simeq -0.324$ e $g'(0.6) \simeq -0.511$, tem-se

$$L = \max_{x \in [0.5, 0.6]} |g'(x)| = |g'(0.6)| \simeq 0.511 < 1 \ .$$

Por conseguinte, é aplicável a seguinte majoração de erro absoluto,

$$|z - x_2| \leq \frac{L}{1 - L} |x_2 - x_1| \simeq 0.0195 \ .$$

Em $I = [0.5, 0.6]$, a função g é continuamente diferenciável e positiva. Como a função g é estritamente decrescente nesse intervalo e

$$g(0.6) \leq g(x) \leq g(0.5) = x_1 < 0.6, \quad \forall x \in I, \quad \text{pois} \quad g(0.6) \simeq 0.509 > 0.5,$$

conclui-se que $g(I) \subset I$. Como $L < 1$, pelo teorema do ponto fixo pode-se garantir que a sucessão (2) converge para o (único) ponto fixo de g em I, qualquer que seja o ponto inicial escolhido nesse intervalo. Logo, a sucessão de ponto fixo, iniciada com $x_0 = 0.5$, é convergente.

(c) O método (1), como já vimos, é o método de Newton e a raiz $z \in (0.5, 0.6)$ é simples. Sabemos que o método tem convergência de ordem 2 (quadrática) local.
Quanto ao método (2), sendo um método do ponto fixo, possui pelo menos convergência de ordem 1. A sua convergência é linear, visto que $g'(z) \neq 0$, onde z é a raiz considerada. Com efeito, vimos que $g'(x) < 0$ em $[0.5, 0.6]$, logo $g'(z) < 0$.

Assim, escolhendo x_0 suficientemente próximo de z, o método (1) converge mais rapidamente para a raiz considerada do que o método (2).

(d) Para se mostrar que o método (2) não é aplicável à raiz negativa da equação considerada, veriquemos que $|g'(z')| > 1$, onde z' é a raiz negativa. Recorrendo ao Teorema de Bolzano, pode concluir-se que $z' \in [-2, -1]$. Calcule-se o valor de g' nos extremos deste intervalo:

$$g'(-2) \simeq 2.93 \quad \text{e} \quad g'(-1) \simeq 1.82 \ .$$

Como já vimos que g' é decrescente, resulta que $|g'(x)| > 1, \forall x \in [-2, -1]$. (Note que não basta verificar que as condições do teorema do ponto fixo não estão satisfeitas em $[-2, -1]$) .

II

1 a) O sistema linear a resolver é da forma

$$\begin{cases} -u + a\,v = -1 \\ u/2 - v = \epsilon \\ u + 3\,w = c. \end{cases}$$

Aplicando o método de Gauss-Seidel, obtém-se

$$\begin{cases} u^{(k+1)} = a\,v^{(k)} + 1 \\ v^{(k+1)} = \frac{1}{2}u^{(k+1)} - \epsilon = \dfrac{a}{2}v^{(k)} + \dfrac{1}{2} - \epsilon \\ w^{(k+1)} = \dfrac{c - u^{(k+1)}}{3} = \dfrac{c - av^{(k)} - 1}{3} \ . \end{cases}$$

1 b) Para verificar que o método de Gauss-Seidel converge, basta mostrar que $\rho(C_{GS}) < 1$, onde C_{GS} é a matriz de iteração do método de Gauss-Seidel para o sistema. Temos

$$C_{GS} = -M^{-1}N = \begin{bmatrix} 0 & a & 0 \\ 0 & a/2 & 0 \\ 0 & -a/3 & 0 \end{bmatrix} \ .$$

Note que a matriz C_{GS} pode obter-se imediatamente a partir das fórmulas iterativas em **1 a)**.

Os valores próprios de C_{GS} são,

$$\lambda_{1,2} = 0 \quad \text{e} \quad \lambda_3 = a/2 \ .$$

Logo, $\rho(C_{GS}) = |a|/2$, pelo que o método de Gauss-Seidel converge se e só se $|a| < 2$.

1 c) Para $a = 0$, a matriz de iteração do método de Jacobi é da forma,

$$C_J = -D^{-1}(L + U) = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ -1/3 & 0 & 0 \end{bmatrix} \ .$$

Pode verificar que $C_J^3 = O$. Assim, qualquer que seja o erro da aproximação inicial $e^{(0)}$, temos

$$e^{(3)} = C^3\, e^{(0)} = 0 \ .$$

Ou seja, a terceira iterada do método tem erro nulo, significando que a solução exacta do sistema é obtida, no máximo, em três iterações (ignorando erros de arredondamento).

2) Tem-se,

$$\begin{aligned} ||A||_\infty &= max\,(1+|a|, 1.5, 4) \\ ||A^{-1}||_\infty &= max\left(\frac{3\,(1+|a|)}{|3-1.5\,a|}, \frac{4.5}{|3-1.5\,a|}, \frac{1+|a|+|1-0.5\,a|}{|3-1.5\,a|}\right) \\ &= \frac{1}{|3-1.5\,a|}\, max\,(3+|3\,a|, 4.5, 1+|a|+|1-0.5\,a|)\,. \end{aligned}$$

Assim,

(i) Se $a \to 2$,

$$||A||_\infty \to 4 \quad \text{e} \quad ||A^{(-1)}||_\infty \to +\infty \ .$$

Por conseguinte, $\lim_{a\to 2} cond_\infty(A) = +\infty$, isto é, a matriz A é mal condicionada.

(ii) Se $|a| \to +\infty$,

$$||A||_\infty \to +\infty, \quad \text{logo a matriz } A \text{ é mal condicionada.}$$

(iii) Se $|a| \to 0$,

$$||A||_\infty = 4 \quad \text{e} \quad ||A||_\infty \to max(3, 4.5, 2)/3 = 4.5/3 \ .$$

Logo $cond_\infty(A) \to 4 \times 4.5/3 = 6$, ou seja, a matriz é bem condicionada.

A.2.14

Exame de 29 de Janeiro de 2014 (Duração: 1h30m) – Parte 2

I

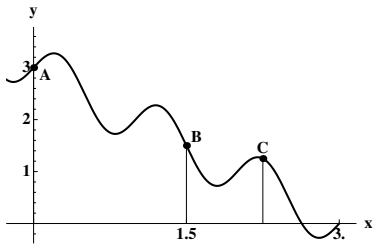
1) Considere os pontos A, B e C, tais que $A = (0, 3)$, $B = (1.5, 1.5)$, $C = (2.25, 1.25)$ e a linha que os une, representada na figura abaixo.

(a) Poderá a referida linha ser o gráfico do polinómio interpolador, com suporte nos pontos A, B, C e só nesses pontos? Justifique sem calcular esse polinómio. **[1.0]**

(b) Usando a fórmula interpoladora de Newton, calcule o polinómio cujo gráfico passa pelos pontos A, B, C. **[1.0]**

2) Pretende-se calcular a função da forma

$$g(x) = a_0 + \frac{a_1}{1 + a_2\, x},$$



que melhor se ajusta aos pontos A, B, C, no sentido dos mínimos quadrados.

(a) Sabendo que o sistema não linear dado a seguir tem, pelo menos, uma solução em $\mathbb{R}^3$, justifique que tal solução nos permite construir a função $g(x)$ pretendida, **[1.0]**

$$\begin{cases} a_0 + a_1 = 3 \\ a_0 + \dfrac{a_1}{1 + 1.5\, a_2} = 1.5 \\ a_0 + \dfrac{a_1}{1 + 2.25\, a_2} = 1.25 \ . \end{cases}$$

[1.0] **(b)** Se utilizar o método de Newton para aproximar a solução do sistema anterior, partindo de uma aproximação inicial tal que $a_0^0 = 1/2$, $a_1^0 = 3$ e $a_2^0 = 1$, qual o sistema de equações lineares que deverá resolver na 1a iteração? (Deduza a matriz e o segundo membro do sistema, não é necessário resolvê-lo).

3) Pretende-se calcular um valor aproximado da área delimitada pela linha curva da figura dada, pelas rectas $x = 0$ e $x = 2.5$, bem como pelo eixo das abcissas.

[1.0] Para o efeito, considere uma regra de quadratura do tipo

$$Q(f) = A_0\, f(0) + A_1\, f(1.5) + A_2\, f(2.5) \ .$$

(a) Escreva um sistema de equações que permita deduzir o valor dos pesos A_0, A_1 e A_2, de modo que $Q(f)$ tenha pelo menos grau 2. (Não é necessário resolver o sistema).

[1.5] **(b)** Sabendo que os pesos da referida regra são $A_0 = 0.5555556$, $A_1 = 1.736110$, e $A_2 = 0.208333$, diga qual é o grau de precisão da regra de quadratura que obteve. Justifique.

II

Considere o problema de valor inicial

$$y'(x) = 2 + \frac{x}{3} + a\, e^{y(x)}, \qquad y(1) = 0.5, \tag{A.1}$$

onde a é um número real. Sejam (A) e (B) dois métodos numéricos para aproximar o problema (1), dados pelas fórmulas,

$$(A) \quad y_{i+1} = y_i + h\left(2 + \frac{x_i}{3} + a\, e^{y_i}\right);$$

$$(B) \quad y_{i+1} = y_i + \frac{h}{2}\left(4 + \frac{x_i + x_{i+1}}{3} + a\left(e^{y_i} + e^{y_i + h\left(2 + \frac{x_i}{3} + a e^{y_i}\right)}\right)\right).$$

(a) Diga, justificando, a que método corresponde cada fórmula. **[1.0]**

(b) Considere a seguinte tabela:

N	Método 1	Método 2
20	4.26771	4.02089
40	4.26944	4.13903

Os resultados apresentados referem-se a valores aproximados de $y(2)$, onde y é a solução exacta do problema. Os valores dispostos em cada coluna foram obtidos pelo mesmo método, usando valores de N distintos, onde $N+1$ é o número de nós utilizados. Sabendo que o valor exacto é $y(2) = 4.26990$, sem reproduzir os cálculos dos valores tabelados diga, justificando, a qual das fórmulas (A) ou (B) corresponde cada um dos métodos a que se refere a tabela. **[1.0]**

(c) No caso de $a = 0$, justifique que o método correspondente à fórmula (B), nos dá o valor exacto da solução do problema (1), para qualquer $x_i = 1 + i\,h$, $i = 0, 1, 2, \cdots$. **[1.5]**

Resolução

I

1 (a) Dados três pontos A, B e C, o polinómio interpolador $p_2(x)$ seria de grau ≤ 2. Assim, p_2, teria no máximo um ponto de máximo ou de mínimo local, o que não acontece na linha figurada. Por isso, o gráfico considerado não pode dizer respeito ao polinómio interpolador.

1 (b). Atendendo a que

$$f[0, 1.5] = \frac{1.5 - 3}{1.5 - 0} = -1$$

$$f[1.5, 2.25] = \frac{1.25 - 1.5}{2.25 - 1.5} \simeq -0.333333$$

$$f[0, 1.5, 2.25] = \frac{f[1.5, 2.25] - f[0, 1.5]}{2.25 - 0} \simeq 0.296296,$$

o polinómio interpolador de Newton, escreve-se

$$\begin{aligned} p_2(x) &= f(0) + f[0, 1.5]\,x + f[0, 1.5, 2.25]\,x\,(x - 1.5) \\ &\simeq 3 - x + 0.296296\,x\,(x - 1.5)\ . \end{aligned}$$

2 (a) Dado que a função g possui três incógnitas, a melhor aproximação de mínimos quadrados será tal que o respectivo gráfico passa pelos 3 pontos dados. Assim, o sistema dado resulta imediatamente das equações $g(0) = 3$, $g(1.5) = 1.5$ e $g(2.25) = 1.25$.

2 (b) Seja

$$F(a_0, a_1, a_2) = (a_0 + a_1 - 3, a_0 + \frac{a_1}{1 + 1.5\, a_2} - 1.5, a_0 + \frac{a_1}{1 + 2.25\, a_2} - 1.25)^T,$$

cuja matriz jacobiana é

$$J_F(a_0, a_1, a_2) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \dfrac{1}{1 + 1.5\, a_2} & -\dfrac{1.5\, a_1}{(1 + 1.5\, a_2)^2} \\ 1 & \dfrac{1}{1 + 2.25\, a_2} & -\dfrac{2.25\, a_1}{(1 + 2.25\, a_2)^2} \end{bmatrix} .$$

Assim,

$$J_F(1/2, 3, 1) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0.4 & 0.12 \\ 1 & 0.31 & 0.64 \end{bmatrix} .$$

e

$$F(1/2, 3, 1) = (0.5, 0.2, 0.17) \ .$$

O sistema linear a resolver, seja $A\, x = b$, tem por matriz $A = J_F(1/2, 3, 1)$ e $b = -F(1/2, 3, 1)$.

3 (a) Para $I(f) = \int_0^{2.5} f(x)dx$, o sistema a resolver tem por equações $Q(1) = I(1)$, $Q(x) = I(x)$ e $Q(x^2) = I(x^2)$, isto é,

$$\begin{cases} A_0 + A_1 + A_2 & = 2.5 \\ \quad 1.5\, A_1 + 2.5\, A_2 & = 2.5^2/2 \\ \quad 1.5^2\, A_1 + 2.5^2\, A_2 & = 2.5^3/3, \end{cases}$$

cuja solução nos dá uma regra exacta para qualquer polinómio de grau ≤ 2, ou seja, de grau pelo menos 2.

3 (b) Seja $f(x) = x^3$. Como

$$Q(f) = 1.736110 \times 1.5^3 + 0.208333 \times 2.5^3 \simeq 9.1$$

e

$$I(f) = 2.5^4/4 \simeq 9.8 \neq Q(f),$$

conclui-se que a regra em causa é de grau 2 de precisão.

II

(a) Dado que

$$y' = f(t, y) = 2 + x/3 + a\, e^y,$$

a fórmula (A) corresponde a $y_{i+1} = y_i + h\, f(t_i, y_i)$, ou seja, ao método de Euler explícito.

O método de Heun é da forma

$$y_{i+1} = y_i + \frac{h}{2}\, (f(x_i, y_i) + f(x_{i+1}, y_i + h\, f(x_i, y_i)) \qquad (*)$$

Para o p.v.i. dado,

$$\begin{aligned} f\left(x_{i+1}, y_i + h\, f(x_i, y_i)\right) &= f\left(x_{i+1}, y_i + h\,(2 + \frac{x_i}{3} + a\, e^{y_i})\right) \\ &= 2 + \frac{x_{i+1}}{3} + a\, e^{\,y_i + h\,(2 + \frac{x_i}{3} + a\, e^{y_i})} . \end{aligned}$$

Substituindo em $(*)$, resulta a fórmula (B).

(b) Sejam h_1 o passo para $N = 20$ e h_2 o passo para $N = 40$, e y_{h_1}, y_{h_2} os respectivos valores tabelados. Os respectivos erros de truncatura são calculados a seguir.

Para o Método 1:

$$\begin{aligned} e_{h_1} &= 4.26990 - 4.26771 = 2.19 \times 10^{-3} \\ e_{h_2} &= 4.26990 - 4.26944 = 4.6 \times 10^{-4} \ . \end{aligned}$$

Para o Método 2:

$$\begin{aligned} e_{h_1} &= 4.26990 - 4.02089 = 0.24901 \\ e_{h_2} &= 4.26990 - 4.13903 = 0.13087 \ . \end{aligned}$$

No Método 2, na passagem de h_1 para h_2, o erro é aproximadamente reduzido a metade. Trata-se, portanto, de um método de primeira ordem de convergência. Quanto ao Método 1, o erro respectivo é aproximadamente reduzido a $1/4$, tratando-se por conseguinte de um método de segunda ordem. Assim, o Método 1 corresponde ao método de Heun, enquanto que o Método 2 diz respeito ao método de Euler explícito.

(c) Para $a = 0$, a equação diferencial é da forma $y'(x) = 2 + x/3$, pelo que a solução do p.v.i. é um polinómio do segundo grau. Atendendo a que o método de Heun pode ser obtido por aplicação da regra dos trapézios, e sabendo que esta regra é exacta para polinómios de grau ≤ 1, conclui-se neste caso que o método de Heun será exacto.

Poderemos chegar à mesma conclusão, notando que o método de Heun é de segunda ordem de convergência. Por conseguinte, o respectivo erro global depende da terceira derivada $y^{(3)}$, derivada esta que no presente caso é nula, daí resultando que o erro deste método será nulo.

Com efeito, para $a = 0$, o método de Heun, com $h = x_{j+1} - x_j$, reduz-se a

$$y_{j+1} = y_j + \frac{h}{2}\left(2 + \frac{x_j}{3} + 2 + \frac{x_{j+1}}{3}\right) = y_j + h\left(2 + \frac{x_j}{3}\right) + \frac{h^2}{6}\,, \quad j = 0, 1, \ldots \,.$$

Por outro lado, o desenvolvimento de Taylor da a solução exacta, em torno de x_j, escreve-se

$$y(x_{j+1}) = y(x_j) + h\, y'(x_j) + \frac{h^2}{2}\, y''(x_j) + \frac{h^3}{6}\, y'''(\xi_j), \quad \xi_j \in (x_j, x_{j+1})\,.$$

Tendo em conta que

$$y'(x) = 2 + \frac{x}{3}\,, \quad y''(x) = \frac{1}{3}\,, \quad y'''(x) = 0\,,$$

obtém-se para a solução exacta:

$$y(x_{j+1}) = y(x_j) + h\left(2 + \frac{x_j}{3}\right) + \frac{h^2}{6}\ .$$

Portanto, a foormula (B) dá-nos o valor exacto, desde que $y_0 = y(x_0)$.

A.2.15

Teste de 7 de Abril de 2014 (Duração: 1h30m)

1) Considere um triângulo rectângulo, tal que d representa o comprimento da hipotenusa e θ um dos seus ângulos internos agudos. O perímetro P do triângulo pode ser calculado através da expressão

$$P = d \times (1 + \sin(\theta) + \cos(\theta)) \qquad (*)$$

[1.0] Admita que θ é aproximado pelo valor $\bar{\theta} = \pi/3$ e seja $\bar{P}$ o valor obtido para o perímetro P. Mostre que o erro relativo de $\bar{P}$ é, aproximadamente, igual a

$$\delta_{P(\bar{\theta})} \simeq \frac{\pi\left(1-\sqrt{3}\right)}{3\left(3+\sqrt{3}\right)}\,\delta_{\bar{\theta}},$$

qualquer que seja o valor $d > 0$ considerado.

2) Suponha que num certo triângulo rectângulo o perímetro é $P = 11$ e a hipotenusa vale $d = 5$.

[1.0] **2 (a)** A partir da fórmula $(*)$, obtenha uma equação do tipo $f(\theta) = 0$. Mostre analiticamente que essa equação tem exactamente duas raízes reais $z_1 < z_2$, no intervalo $[0, \pi/2]$ e, para cada raiz, determine um intervalo que a contenha.

[2.0] **2 (b)** Considere o método iterativo

$$\begin{aligned} &\theta_0 = 0.75 \\ &\theta_{k+1} = \theta_k + 0.2\,(6 - 5\,\cos(\theta_k) - 5\,\sin(\theta_k))\,, \quad k = 0, 1, \ldots\ . \end{aligned}$$

Prove que o método converge para uma das raízes referidas na alínea anterior. Determine a ordem de convergência do método.

[1.5] **2 (c)** Escreva a fórmula iterativa do método de Newton aplicado à função f que definiu na alínea **2**(**a**). Poderá usar como aproximação inicial $\theta_0 = 0.8$, caso aplique esse método para aproximar a maior raiz z_2? Justifique.

[1.0] **2 (d)** Fazendo $\theta_0 = 0$, efectue duas iterações do método de Newton. Sendo um método de convergência supralinear, use a fórmula $e_k = z_1 - \theta_k \simeq \theta_{k+1} - \theta_k$ para obter uma estimativa do erro absoluto da iterada θ_2 em relação a z_1 .

3) Considere o sistema linear $A\,x = b$, de solução $x = (1, 1)^T$, sendo

$$A = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 5 \\ 2 \end{bmatrix}.$$

[1.5] **3 (a)** Diga, justificando, se o método de Jacobi é convergente para a solução do sistema, considerando a aproximação inicial $x^{(0)} = (-5, 5)^T$.

[1.0] **3 (b)** Considere a norma vectorial $||\,.\,||_1$. Partindo de $x^{(0)} = (0, 0)^T$, calcule um majorante para $||x - x^{(2)}||_1$, onde $x^{(2)}$ é a segunda iterada do método de Gauss-Seidel.

[1.0] **4)** Considere a matriz de entradas reais,

$$H = \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix},$$

onde $c \neq 1$. Calcule $cond_\infty(H)$, isto é, o número de condição da matriz H na norma ∞. O que pode dizer sobre o condicionamento dum sistema da forma $H\,x = v$, quando c se aproxima de 1?

Resolução

1(a) Dado que

$$\delta_{P(\bar\theta)} \simeq \frac{\bar\theta \frac{dP}{d\theta}(\bar\theta)}{P(\bar\theta)}\,\delta_{\bar\theta} = \frac{\bar\theta\,\big(\cos(\bar\theta) - \sin(\bar\theta)\big)}{\big(1 + \sin(\bar\theta) + \cos(\bar\theta)\big)}\,\delta_{\bar\theta},$$

para $\bar\theta = \pi/3$ resulta a expressão dada.

2(a) Para $P = 11$ e $d = 5$, tem-se

$$f(\theta) = 11 - 5\ (1 + \sin(\theta) + \cos(\theta)) = 0, \quad \text{sendo} \quad f \in C^\infty(\mathbb{R})\ .$$

No intervalo $I = [0, \pi/2]$, considerem-se os pontos $\theta = 0$, $\theta = \pi/4$ e $\theta = \pi/2$. Como

$$\begin{array}{l} f(0) = 1 > 0 \\ f(\pi/4) \simeq -1.1 \\ f(\pi/2) = 1, \end{array}$$

conclui-se que em $(0, \pi/4)$ existe pelo menos um zero de f e de igual modo em $(\pi/4, \pi/2)$. Dado que

$$f'(\theta) = -5\,(\cos(\theta) - \sin(\theta)),$$

tem-se que $f'(\theta) = 0$ e $\theta \in (0, \pi/2)$ se e só se $\theta = \pi/4$. A função f' é negativa em $(0, \pi/4)$ e positiva em $(\pi/4, \pi/2)$ pelo que é único o zero de f existente em cada um desses subintervalos.

2(b) É fácil concluir que as raizes da equação $5\,(1 + \sin(\theta) + \cos(\theta)) = 11$ são pontos fixos da função iteradora $g(\theta) = \theta + 0.2\,(6 - 5\,\sin(\theta) - 5\,\cos(\theta))$.
Viu-se na alínea anterior que $z_1 \in [0, \pi/4]$ e $z_2 \in [\pi/4, \pi/2]$. Como g' é crescente e $g'(\pi/4) = 1$, conclui-se respectivamente que z_1 é um ponto fixo atractor e z_2 é um ponto fixo repulsor para a função g. Logo, a sucessão, caso convirja, só pode convergir para z_1 . Aplicando o teorema do ponto fixo em $I = [0, 0.75]$ conclui-se que o método é convergente neste intervalo.
A ordem de convergência é 1, pois $1 > g'(z_1) > 0$.

2(c) O método é da forma

$$\theta_{k+1} = \theta_k + \frac{6 - 5\ (\sin(\theta_k) + \cos(\theta_k))}{5\ (\cos(\theta_k) - \sin(\theta_k))}, \quad k = 0, 1, \ldots, \quad \text{com} \quad \theta_0 = 0.8 \quad (**)$$

Para $x_0 = 0.8$ a primeira iterada é $x_1 \simeq 11.1665$, a qual está fora do intervalo pretendido. Pode verificar-se que, por exemplo, no intervalo $[11, 13.2]$ estão reunidas as condições suficientes para a convergência do método de Newton nesse intervalo. Assim,

a sucessão $(**)$ convergirá para a raiz existente nesse intervalo e não para a raiz z_2 em causa.

2(d)

$$\begin{aligned} x_0 &= 0 \\ x_1 &= x_0 - f(x_0)/f'(x_0) = 1/5 = 0.2 \\ x_2 &= x_1 - f(x_1)/f'(x_1) \simeq 0.227212908 \\ x_3 &= x_2 - f(x_2)/f'(x_2) \simeq 0.227799061 \ . \end{aligned}$$

Assim,

$$z - x_2 \simeq x_3 - x_2 \simeq 0.00059 \ .$$

3(a) As fórmulas computacionais do método são da forma

$$\begin{cases} x_1^{(k+1)} = 5/4 - x_2^{(k)}/4 \\ x_2^{(k+1)} = 2 - x_1^{(k)}, \qquad k = 0, 1, \ldots \end{cases}$$

isto é,

$$x^{(k+1)} = \begin{bmatrix} 0 & -1/4 \\ -1 & 0 \end{bmatrix} x^{(k)} + \begin{bmatrix} 5/4 \\ 2 \end{bmatrix} = C_j\, x^{(k)} + d, \quad k = 0, 1, \ldots \ .$$

A equação característica de C_J é $\lambda^2 - 1/4 = 0$. Por conseguinte, o raio espectral é $\rho(C_J) = 1/2 < 1$. Assim, o método é convergente para a solução do sistema, qualquer que seja a aproximação inicial $x^{(0)}$.

3(b) As fórmulas computacionais do método escrevem-se

$$\begin{cases} x_1^{(k+1)} = (5 - x_2^{(k)})/4 \\ x_2^{(k+1)} = 2 - x_1^{(k+1)} = (3 + x_2^{(k)})/4, \quad k = 0, 1, \ldots . \end{cases}$$

Logo

$$C_{GS} = \begin{bmatrix} 0 & -1/4 \\ 0 & 1/4 \end{bmatrix} \Longrightarrow ||C_{GS}||_1 = 1/2 \ .$$

As duas primeiras iteradas são:

$$\begin{aligned} x^{(1)} &= (5/4, 3/4)^T \\ x^{(2)} &= (17/16, 15/16)^T \Longrightarrow ||x - x^{(2)}||_1 = ||(-3/16, 3/16)||_1 = 6/16 = 3/8 \ . \end{aligned}$$

Uma majoração do erro de $x^{(2)}$ pode ser obtida através da expressão

$$||x - x^{(2)}||_1 \leq \frac{||C_{GS}||_1}{1 - ||C_{GS}||_1} ||x^{(2)} - x^{(1)}||_1 = 3/8 = 0.375 \ .$$

4)

$$H^{-1} = \frac{1}{1 - c^2} \begin{bmatrix} 1 & -c \\ -c & 1 \end{bmatrix}.$$

Assim, $||H||_\infty = 1 + |c|$ e $||H^{-1}||_\infty = \dfrac{1+|c|}{|1-c^2|}$. Por conseguinte,

$$cond_\infty(H) = \frac{(1+|c|)^2}{|1-c^2|} \ .$$

O número de condição poderá tomar valores muito elevados quando c for próximo de 1, o que significa que nesse caso a matriz será mal condicionada.

A.2.16

Exame/teste de recuperação, 03/07/14, Parte 1 (Duração: 1h30m)

1. (a) Prove que a sucessão definida por **[1.5]**

$$x_0 = 1, \quad x_{n+1} = \sqrt{\frac{10 - x_n^3}{4}} \quad = g_1(x_n)$$

converge para um número $\alpha \in [1, 1.5]$.

(b) Sabe-se que a sucessão **[1.5]**

$$y_0 = 1, \quad y_{n+1} = \sqrt{\frac{10}{4 + y_n}} \quad = g_2(y_n)$$

converge <u>linearmente</u> para α e que $|g_2'(y)| \leq 0.15$, para todo o $y \in [1, 1.5]$. Verifique, sem calcular x_3 e y_3, que se tem:

$$|\alpha - x_3| \geq |\alpha - y_3| \ .$$

Diga, justificando, qual das sucessões, $(x_n)_{n\geq 0}$ ou $(y_n)_{n\geq 0}$, converge mais rapidamente para α .

(c)-i Mostre que α é raiz da equação $x^3 + 4x^2 - 10 = 0$. Admita que o método de Newton, aplicado a esta equação, converge para α tomando $z_0 = 1$ para iterada inicial. Calcule a iterada z_2 e obtenha um majorante para o erro absoluto de z_2 . **[1.5]**

(c)-ii Determine a ordem de convergência do método de Newton considerado na alínea anterior e compare-a com a dos 2 métodos estudados em **1.a)-b)**. **[1.0]**

2. Considere o seguinte sistema de equações lineares:

$$\begin{cases} 5x_1 + 2x_2 & = b_1 \\ 2x_{i-1} + 5x_i + 2x_{i+1} & = b_i, \qquad i = 2, \ldots, n-1 \\ 2x_{n-1} + 5x_n & = b_n \end{cases} \tag{A.2}$$

onde $b_i \in R$, $n \geq 2$.

(a) Justifique que, para qualquer $n \geq 2$, este sistema tem uma única solução, $x = (x_1, x_2, \cdots, x_n)^T$, para cada $b = (b_1, b_2, \cdots, b_n)^T \in R^n$. Mostre ainda que o método **[1.0]**

de Gauss-Seidel aplicado ao sistema (A.2) converge, qualquer que seja $n \geq 2$, independentemente da iterada inicial.

[1.5] **(b)-i** No caso de $n = 3$ mostre que as iteradas do método satisfazem a desigualdade

$$\|x^{(k)} - x\|_\infty \leq \frac{14}{11} \ \|x^{(k)} - x^{(k-1)}\|_\infty \ .$$

[0.5] **(b)-ii** Ainda no caso de $n = 3$, sendo $x^{(0)} = (1, 1, 1)^T$ e $b = (0, 0, 0)^T$, obtenha a primeira iterada, $x^{(1)}$, do método de Gauss-Seidel.

[1.5] **(c)** Sabendo que os valores próprios λ_i da matriz A do sistema (A.2) satisfazem

$$1 < \lambda_i < 9, \qquad i = 1, ..., n$$

diga se o sistema é bem condicionado, para qualquer valor de n, e indique um majorante de $cond_2(A)$.

Resolução

1- (a) Pretende-se provar que a sucessão $x_0 = 1, \quad x_{n+1} = \sqrt{\dfrac{10 - x_n^3}{4}} \quad = g_1(x_n)$ converge para um número $\alpha \in [1, 1.5]$.
Sabemos que, sendo g_1 contínua em $I = [1, 1.5]$, se a sucessão gerada por g_1 convergir, o seu limite será um ponto fixo de g_1. Verifiquemos as condições do teorema do ponto fixo em I :

i) As funções g_1 e $g_1'(x) = -\dfrac{3x^2}{4\sqrt{10 - x^3}}$ são contínuas em I (o denominador de g_1' anula-se para $x \simeq 2.15443$) .

ii) Tem-se $g_1(1.0) = 1.5 \in I, \quad g_1(1.5) = 1.28695 \in I$, e como g_1 decrescente ($g_1' < 0$), então $g_1(x) \in I, \forall\, x \in I$.

iii) Para provar que $\max_{x \in I} |g_1'(x)| = L$, com $0 < L < 1$, calculemos os valores

$$g_1'(1) = -0.25$$
$$g_1'(1.5) = -0.6556,$$

cujos módulos são < 1 . Para avaliar o que se passa nos restantes pontos, investiguemos a monotonia de g_1'. Como $g_1''(x) = -(\dfrac{9\,x^4}{8\,(10 - x^3)^{3/2}} + \dfrac{3\,x}{2\,\sqrt{10 - x^3}})$, podemos concluir que g_1' é monótona decrescente (e negativa em I.) Logo $|g_1'|$ é monótona crescente (e obviamente positiva), pelo que o seu máximo é atingido num dos extremos (no direito). Resulta,

$$L = \max\{0.25, 0.6556\} = 0.6556 < 1 \ .$$

Das condições (i)-(iii), conclui-se que a sucessão gerada por g_1 converge para o único ponto fixo de g_1 pertencente ao intervalo I, qualquer que seja $x_0 \in I$ e, em particular, se $x_0 = 1$. Designamos esse ponto fixo por α .

1-(b) Para a sucessão $y_0 = 1, \quad y_{n+1} = g_1(y_n)$, tem-se

$$|y_1 - \alpha| = |g_2'(\xi_1)| \; |y_0 - \alpha| \leq \max_{[1,1.5]} |g_2'(x)| \; |e_0| \leq 0.15 \; |e_0|,$$

e $|y_2 - \alpha| = |g_2'(\xi_2)| \; |y_1 - \alpha| \leq (0.15)^2 \; |e_0|$. Analogamente, $|y_3 - \alpha| \leq (0.15)^3 \; |e_0|$, e, em geral,

$$|y_k - \alpha| \leq (0.15)^k |e_0| \; .$$

Consideremos agora a sucessão $x_0 = 1, \quad x_{n+1} = g_1(x_n)$.

Note-se que já obtivémos na alínea anterior $0.25 \; \leq |g_1'(x)| \leq 0.6556, \quad x \in [1., 1.5]$. Sendo $x_0 = y_0 = 1$, então $|y_0 - \alpha| = |x_0 - \alpha| = |e_0|$. Por um processo análogo ao utilizado para (y_n), são válidas as desigualdades,

$$|x_1 - \alpha| = |g_1'(\xi_1)| \; |x_0 - \alpha| \leq \max_{[1,1.5]} |g_1'(x)| \leq 0.6556 \; |e_0|,$$

$$|x_1 - \alpha| = |g_1'(\xi_1)| \; |x_0 - \alpha| \geq \min_{[1,1.5]} |g_1'(x)| \geq 0.25 \; |e_0| \; .$$

Obtém-se, em geral,

$$(0.25)^k \; |e_0| \leq \; |x_k - \alpha| \leq (0.06556)^k \; |e_0|, \quad \text{para} \quad k = 1, 2 \ldots \; .$$

Combinando $(0.25)^3 |e_0| \leq |x_3 - \alpha|$ e $|y_3 - \alpha| \leq (0.15)^3 \; |e_0|$, fica provada a desigualdade $|y_3 - \alpha| \leq |x_3 - \alpha|$. Na verdade, é válida a desigualdade estrita

$$|y_k - \alpha| < |x_k - \alpha|, \quad k \geq 1 \; . \tag{A.3}$$

Atendendo a (A.3), podemos concluir que a sucessão (y_n) converge mais rapidamente.

1.(c)-i Sabemos que α é o único ponto fixo de g_1 em $I = [1, 1.5]$. Então,

$$\alpha = g_1(\alpha) \Longleftrightarrow \alpha = \sqrt{\frac{10 - \alpha^3}{4}} \Longrightarrow \alpha^2 = (10 - \alpha^3)/4 \; .$$

Donde, α satisfaz a equação $\alpha^3 + 4\alpha^2 - 10 = 0$. Com $z_0 = 1$, vem

$$\begin{aligned} z_1 &= z_0 - f(z_0)/f'(z_0) = 1.45455 \\ z_2 &= z_1 - f(z_1)/f'(z_1) = 1.3689 \; . \end{aligned}$$

Um majorante para $|\alpha - z_2|$ obtém-se da fórmula

$$|\alpha - z_k| \leq \frac{1}{\mathbb{K}} \left(\mathbb{K}|\alpha - z_0|\right)^{2^k},$$

fazendo $k = 2$. Assim,

$$|\alpha - z_2| \leq \frac{1}{\mathbb{K}} \left(\mathbb{K}|\alpha - z_0|\right)^4 \leq 0.028838,$$

onde se tomou $|\alpha - z_0| \leq 0.5$ e

$$\mathbb{K} = \frac{\max|f''|}{2\min|f'|} = \frac{17}{22} \simeq 0.7727, \qquad x \in I \ .$$

1.(c)-ii Sabemos que o método de Newton tem ordem pelo menos quadrática. Por outro lado, as iteradas do método de Newton satisfazem a igualdade,

$$\lim_{m\to\infty} \frac{|\alpha - z_{m+1}|}{(\alpha - z_m)^2} = \frac{|f''(\alpha)|}{2\,|f'(\alpha)|} \ . \tag{A.4}$$

Dado que $f''(x) = 8 + 6\,x$ não se anula no intervalo $I = [1, 1.5]$ a que pertence α, resulta $f''(\alpha) \neq 0$. O limite (A.4) é diferente de zero e, atendendo à definição de ordem de convergência, a ordem do método é exactamente dois.

Comparemos com a ordem das outras sucessões anteriormente referidas. Sabe-se que (y_n) converge linearmente para α . Por outro lado, já obtivémos na alínea **1-a)** que $0.25 \ \leq |g_1'(x)| \leq 0.6556, \quad x \in [1., 1.5]$. Então $g_1'(\alpha) \neq 0$, pelo que a sucessão (x_n) também é de convergência linear.

2.(a) A forma geral da matriz do sistema considerado é

$$A = \begin{bmatrix} 5 & 2 & 0 & \dots & 0 \\ 2 & 5 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 2 & 5 \end{bmatrix} \ .$$

Esta matriz tem a diagonal estritamente dominante por linhas (e por colunas) – já que em módulo a entrada da diagonal principal é 5 (em todas as linhas), enquanto que a soma dos módulos das entradas não diagonais é não superior a 4. Por conseguinte, a matriz é não singular e o sistema tem solução única. Além disso, o método de Gauss-Seidel converge, quando aplicado a este sistema.

2. (b)-i A estimativa do erro das iteradas do método de Gauss-Seidel, no caso geral, é

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{C}\|}{1 - \|\mathbf{C}\|}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|,$$

onde C representa a matriz de iteração do método. Para $n = 3$, temos

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix},$$

pelo que

$$C = -(L + D)^{-1}U = \begin{bmatrix} 0 & -2/5 & 0 \\ 0 & 4/25 & -2/5 \\ 0 & -8/125 & 4/25 \end{bmatrix}.$$

Por conseguinte,

$$\|C\|_\infty = \max(2/5, 14/25, 28/125) = 14/25 \ .$$

Finalmente, uma vez que

$$\frac{\|C\|_\infty}{1-\|C\|_\infty} = 14/11,$$

obtém-se a desigualdade que se pretende demonstrar.

2.(b)-ii Aplicando a fórmula do método de Gauss-Seidel, tem-se:

$$\begin{aligned} x_1^{(1)} &= -\frac{2\,x_2^{(0)}}{5} = -2/5 \\ x_2^{(1)} &= -\frac{2\,x_1^{(1)} + 2\,x_3^{(0)}}{5} = -6/25 \\ x_3^{(1)} &= -\frac{2\,x_2^{(1)}}{5} = 12/125 \ . \end{aligned}$$

2.(c) Sabendo que os valores próprios da matriz do sistema satisfazem $1 < \lambda_i < 9$, podemos imediatamente concluir que $\rho(A) < 9$ (onde $\rho(A)$ representa o raio espectral de A). Além disso, como a matriz é simétrica, temos $\|A\|_2 = \rho(A) < 9$.
Em relação à inversa de A, sabemos que os seus valores próprios são os inversos de λ_i, e dos dados do problema concluímos,

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \max\left(1/\lambda_i\right) < 1 \ .$$

Finalmente, pela definição de número de condição,

$$\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 < 9 \ .$$

Visto que $cond_2(A)$ não é um número muito elevado, conclui-se que a matriz é bem condicionada, independentemente da sua dimensão.

A.2.17

Exame de 3 de Julho de 2014, Parte 2

1 (a) Dados três nós distintos x_1, x_2 e x_3 e uma tabela $\{x_i, f(x_i)\}$, $i = 1, 2, 3$ (onde f é uma função genérica), considere a base de Newton associada aos nós $\{1, x - x_1, (x - x_1)\,(x - x_2)\}$. Mostre que o polinómio interpolador dos valores tabelados, representado na base referida, pode ser obtido resolvendo um certo sistema linear de equações, o qual deve determinar. Diga, justificando, se tal sistema pode ou não ter mais do que uma solução. **[1.5]**

1 (b) Considere a tabela

x_i	0.8	1.6	2.4
$f(x_i)$	0.7	1.0	0.7

Determine o respectivo polinómio interpolador de Newton. **[1.0]**

1 (c) Calcule uma aproximação de $f(1)$ através do polinómio interpolador da tabela. Admita que a função f é continuamente diferenciável até à ordem que pretender e que $f(x) - 3/x$ é um polinómio de grau ≤ 2. Obtenha um majorante do erro absoluto da referida aproximação. Justifique. **[1.5]**

1 (d) Usando a definição de melhor aproximação de mínimos quadrados, e sem efectuar cálculos, diga qual é a melhor aproximação da tabela dada mediante funções aproximantes do tipo $h(x) = a + b\,x + c\,x^2$. **[1.0]**

2 (a) Sendo $\phi(x) = \cos(\pi x/2)$, calcule um valor aproximado de $\displaystyle\int_0^1 \sqrt{1 + [\phi(x)]^2}dx$, por aplicação da regra de Simpson com 4 subintervalos. **[1.0]**

Para aproximar um integral $I(f) = \int_a^b f(x)dx$, adoptou-se uma fórmula de quadratura da forma $Q(f) = A\,f(B)$, com $A, B \in \mathbb{R}$ e f integrável.

2 (b) Diga o que entende por grau de precisão da regra $Q(f)$. **[1.0]**

2 (c) Determine as constantes A e B, de modo que a regra tenha grau 1. Conhece outra regra de quadratura de grau 1? Justifique. **[1.5]**

3) Considere o problema de valor inicial

$$\begin{array}{ll} y'(x) = \sin(x)\,\cos(y(x)), & x \in [0,1] \\ y(0) = 2\ . & \end{array}$$

3 a) Para um dado passo h, escreva a equação às diferenças do correspondente método de Taylor de segunda ordem. Justifique. **[1.0]**

3 b) Fazendo $h = 0.1$ obtenha uma aproximação de $y(0.1)$, por aplicação do método anteriormente referido. **[0.5]**

Resolução

1 (a) Seja $p_2(x) = a_0 + a_1\,(x - x_1) + a_2\,(x - x_1)\,(x - x_2)$ o polinómio interpolador de Newton. Os seus coeficientes podem ser determinados considerando as condições interpolatórias $p_2(x_1) = f(x_1)$, $p_2(x_2) = f(x_2)$ e $p_2(x_3) = f(x_3)$, isto é,

$$\left\{\begin{array}{ll} a_0 & = f(x_1) \\ a_0 + (x_2 - x_1)\,a_1 & = f(x_2) \\ a_0 + (x_3 - x_1)\,a_1 + (x_3 - x_1)\,(x_3 - x_2) & = f(x_2)\ . \end{array}\right.$$

Como os nós são distintos e o sistema anterior possui matriz triangular, o seu determinante vale $(x_2 - x_1)\,(x_3 - x_1)\,(x_3 - x_2) \neq 0$, pelo que existe solução única do sistema.

1 (b)

x_i	f_i	$f[..]$	$f[...]$
0.8	0.7		
		0.375	
1.6	1.0		-0.46875
		-0.375	
2.4	0.7		

$p_2(x) = 0.7 + 0.375\,(x - 0.8) - 0.46875\,(x - 0.8)\,(x - 1.6)$.

1 (c) $p_2(1) = 0.83125 \simeq f(1)$. Dado que

$$f(1) - p_2(1) = \frac{f^{(3)}(\xi)}{3!}(1 - 0.8)\,(1 - 1.6)\,(1 - 2.4), \qquad \xi \in (0.8, 2.4),$$

e como $f(x) - 3/x = q(x)$, e $q \in \mathcal{P}_2 \Longrightarrow f^{(3)}(x) + 18/x^4 = 0$, obtém-se a majoração

$$|f(1) - p_2(1)| \leq \frac{18}{0.8^4 \times 6} \times 0.2 \times 0.6 \times 1.4 \simeq 1.23 \ .$$

1 (d) Visto que o menor valor possível de $\sum_{i=1}^{3}(f(x_i) - (a + b\,x + c\,x^2))^2$ é zero, e este valor ocorre quando a função h coincide com o polinómio interpolador da tabela, conclui-se que este polinómio é a melhor aproximação pretendida, sendo $c = -0.46875$, $b = 1.4$ e $a = -0.2$.

2 (a) Seja $f(x) = \sqrt{1 + \phi(x)^2}$. Para a regra de Simpson, com passo $h = 1/4$, obtém-se

$$\begin{aligned} S(f) \ &= \frac{h}{3}\,[f(0) + f(1) + 4\,(f(1/4) + f(3/4)) + 2\,f(1/2)] = \\ &= \frac{1}{12}\,[1 + 2/\sqrt{(2)} + 4\,\left(\sqrt{1 + \cos(\pi/8)^2} + \sqrt{1 + \sin(\pi/8)^2}\right) + \\ &\quad +2\,\sqrt{1 + \cos^2(\pi/4)}] \simeq 1.21603 \ . \end{aligned}$$

2 (b) A regra possui grau de precisão $k\,(k \geq 0)$ se e só se é exacta para qualquer polinómio de grau menor ou igual a k, e existe algum polinómio de grau $k + 1$ para o qual não é exacta.

2 (c) A regra é de grau 1 se for exacta para 1 e x, isto é,

$$A = \int_a^b dx = b - a \quad \text{e} \quad A\,B = \int_a^b x\,dx = (b^2 - a^2)/2 \Longleftrightarrow B = (a + b)/2 \ .$$

Ou seja, $Q(f) = (b - a)\,f((a + b)/2)$ é a regra do ponto médio.

3 (a) Para $y'(x) = f(x, y(x)) = \sin(x)\,\cos(y(x))$, tem-se $y''(x) = \cos(x)\,\cos(y(x)) - \sin(x)\,\sin(y(x))\,y'(x)$, ou seja,

$$\begin{aligned} y''(x) \ &= \cos(x)\,\cos(y(x)) - \sin^2(x)\,\sin(y(x)\,\cos(y(x))) = \\ &= \cos(y(x))\,\left[\cos(x) - \sin^2(x)\,\sin(y(x))\right] \ . \end{aligned}$$

Atendendo a que

$$y(x+h) \simeq y(x) + h\,y'(x) + h^2/2\,y''(x),$$

o método de Taylor de segunda ordem tem a forma

$$\begin{array}{ll} y_0 & = 2 \\ y_{i+1} & = y_i + h\,[\sin(x_i)\,\cos(y_i)] + h^2/2\,\left\{\cos(y_i)\,[\cos(x_i) - \sin^2(x_i)\,\sin(y_i)]\right\}, \quad i = 0, 1, \ldots \end{array}$$

3 (b) Para $h = 0.1$, $x_0 = 0$ e $y_0 = 2$, obtém-se

$$y(0.1) \simeq y_1 = y_0 + h^2/2\,\cos(y_0) = 2 + 0.1^2/2\,\cos(2) \simeq 1.99792\ .$$

A.2.18

Teste de 13 de Novembro de 2014 (Duração: 1h30m)

[1.0] **1)** Dada a função $f(x) = \ln(x)$, determine uma estimativa para o erro relativo que se comete no cálculo de $f(1.01)$, quando em vez de $x = 1.01$ é utilizada a aproximação $\tilde{x} = 1.009$.

2) Considere a equação

$$f(x) \equiv x\,e^{-x} - e^{-2} = 0 \qquad (1)$$

[1.0] **(a)** Mostre que a equação (1) tem uma única raiz z no intervalo $[0, 1]$.

[1.5] **(b)** Seja $g(x) = e^{x-2}$. Verifique que z é ponto fixo de g. Mostre que a sucessão $x_{m+1} = g(x_m)$ converge para a raiz z da equação (1), qualquer que seja a aproximação inicial x_0 escolhida no intervalo $[0, 1]$.

[1.0] **(c)** Tomando $x_0 = 1$, determine uma estimativa para o número de iterações necessárias para garantir uma aproximação x_k de z, com erro absoluto inferior a 10^{-6} .

[1.5] **(d)** Considere a sucessão $y_{n+1} = G(y_n)$, para $n \geq 0$, definida por,

$$G(x) = \frac{g(x) - z\,x}{1 - z},$$

onde g é a função dada em (b) e z é a raiz da equação (1). Mostre que a sucessão $\{y_n\}$ converge para z, se partir de y_0 suficientemente próximo de z. Indique a respectiva ordem de convergência e compare-a com a da sucessão $\{x_m\}$.

3) Considere o sistema linear $A\,w = b$, onde

$$A = \begin{bmatrix} 3 & 0 & c \\ 0 & 3 & 2 \\ c & 2 & 5 \end{bmatrix} \quad \text{e} \quad b = \left[\sqrt{3} - 4, -8, -2\right]^T \ .$$

[1.5] **(a)** Determine <u>todos</u> os valores de c para os quais o método se Jacobi aplicado aos sistema é convergente, qualquer que sela a aproximação inicial $w^{(0)}$ de $\mathbb{R}^3$ que considere.

[1.0] **(b)** Faça $c = 2$. Tomando a aproximação inicial $w^{(0)} = [1, 0, 2]^T$, calcule a primeira iterada $w^{(1)}$ do método de Jacobi. Utilize-a para obter um majorante de $||w - w^{(2)}||_\infty$.

[0.5] **(c)** Ainda com $c = 2$, sabendo que $||A||_1 = 9\,||A^{-1}||_\infty$, calcule o número de condição da matriz A, na norma $||\ .\ ||_\infty$. Diga para que serve, justificando.

4) Pretende-se aplicar o método de Newton ao sistema de equações não lineares, **[1.0]**

$$\begin{array}{ll} 3\,x_1 + x_3^2 & = \sqrt{3} \\ 3\,(x_2 + 1) + x_3^2 & = 11 \\ 2\,x_1 + x_2\,(x_3 + 1) & = 10\ . \end{array}$$

Tomando como aproximação inicial $x^{(0)} = [1, 5, 1]^T$, mostre que o sistema linear a ser resolvido para se obter $x^{(1)}$ é o sistema $A\,w = b$ considerado na questão $3(b)$. Em seguida, calcule $x^{(1)}$, utilizando a iterada $w^{(1)}$ obtida em $3(b)$ para aproximar w.

Resolução

1) Como $x - \tilde{x} = 0.001$, tem-se $\delta_{\tilde{x}} = (x - \tilde{x})/x = 0.00099$. Atendendo a que $f(1.01) \simeq 0.0099503309$ e $f(1.009) \simeq 0.0089597414$, o erro relativo propagado à função pode obter-se mediante a expressão

$$\delta_{f(\tilde{x})} = \frac{f(x) - f(\tilde{x})}{f(x)} \simeq 0.0996 \simeq 10\,\% \ .$$

Uma estimativa pode ser calculada através de

$$\delta_{f(\tilde{x})} = \frac{\tilde{x}\,f'(\tilde{x})}{f(\tilde{x})}\,\delta_{\tilde{x}} = \frac{1}{f(\tilde{x})} \simeq 0.111 \simeq 11\% \ .$$

ou ainda,

$$\delta_{f(\tilde{x})} = \frac{x\,f'(x)}{f(x)}\,\delta_{\tilde{x}} = \frac{1}{f(x)} \simeq 0.0995 \simeq 10\% \ .$$

2(a) A função $f(x) = x\,e^{-x} - e^{-2}$ é de classe $C^\infty(I)$. Dado que $f(0) < 0$ e $f(1) > 0$, existe pelo menos um zero de f em $(0, 1)$. Atendendo a que $f'(x) = e^{-x} - x\,e^{-x} = e^{-x}\,(1 - x) \geq 0, \quad \forall x \in I$, a função é crescente no intervalo, pelo que existe no máximo um zero nesse intervalo. Por conseguinte existe um só valor z em $(0, 1)$, tal que $f(z) = 0$.

2(b) No intervalo I considerado, a equação $g(x) = x$ satisfaz as equivalências

$$e^{x-2} = x \iff e^{-2} = x\,e^{-x} \iff f(x) = 0\ .$$

Assim, se z é zero de f, tem-se $g(z) = z$, isto é, z é ponto fixo de g. Dado que a função g é positiva e $g \in C^1(I)$, como $g'(x) = g(x) > 0$, resulta que g é estritamente crescente no intervalo, logo

$$0 < g(0) \leq g(x) \leq g(1) < 1\ .$$

Como

$$max_{x\in I}|g'(x)| = g(1) = e^{-1} = L < 1,$$

pelo teorema do ponto fixo podemos garantir convergência do processo iterativo, qualquer que seja o valor inicial $x_0 \in I$ escolhido.

2(c) Seja $\epsilon = 10^{-6}$. Como $|z - x_k| \leq L^k\,|z - x_0|$ e $|z - x_0| < 1$, vem $|z - x_k| < L^k$. Assim,

$$L^k < \epsilon \Longrightarrow k > \log(\epsilon)/\log(L) \simeq 13.8\ .$$

Por conseguinte, efectuando $k = 14$ iterações podemos garantir que $|e_{14}| = |z - x_{14}| < \epsilon$.

2(d) No intervalo I a função iteradora $G \in C^2$. Como $z \neq 1$ e

$$G(z) = \frac{g(z) - z^2}{1 - z} = \frac{z\,(1 - z)}{1 - z} = z,$$

conclui-se que z é ponto fixo de G . Além disso,

$$G'(x) = \frac{g'(x) - z}{1 - z} \quad \text{e} \quad g'(x) = g(x),$$

logo

$$G'(z) = \frac{g(z) - z}{1 - z} = 0 \quad (\text{pois } z \text{ é ponto fixo de } g)\ .$$

Como

$$G''(x) = \frac{g''(x)}{1 - z} = \frac{g(x)}{1 - z} \Longrightarrow G''(z) = \frac{z}{1 - z} \neq 0,$$

conclui-se que o método $y_{k+1} = G(y_k)$ converge localmente para z e a sua ordem de convergência é 2. Pelo contrário, a sucessão $x_{m+1} = g(x_m)$ possui ordem 1 de convergência visto que $g'(z) \neq 0$.

3(a) Dado que

$$C_J = -D^{-1}\,(L + U) = \begin{bmatrix} 0 & 0 & -c/3 \\ 0 & 0 & -2/3 \\ -c/5 & -2/5 & 0 \end{bmatrix},$$

sabemos que é condição necessária e suficiente de convergência deste método iterativo que o raio espectral da matriz C_J seja inferior a 1. Ora,

$$det(C_J - \lambda\, I) = \lambda\left(\lambda^2 - 4/15\right) + c/5\,(-c\,\lambda/3) = 0,$$

isto é,

$$\lambda\,(\lambda^2 - (4 + c^2)/15)) = 0\ .$$

Assim , o espectro de C_J é $\{0, -\sqrt{(4 + c^2)/15}, +\sqrt{(4 + c^2)/15}\}$, pelo que

$$\rho(C_J) < 1 \Leftrightarrow c^2 < 11 \Leftrightarrow |c| < \sqrt{11}\ .$$

3(b) Dado $w^{(0)} = (1, 0, 2)^T$, as fórmulas computacionais do método escrevem-se,

$$\begin{array}{rcl} w_1^{(k+1)} & = & \dfrac{\sqrt{3} - 4 - 2\, w_3^{(k)}}{3} \\ w_2^{(k+1)} & = & \dfrac{-8 - 2\, w_3^{(k)}}{3} \\ w_3^{(k+1)} & = & \dfrac{-2 - 2\, w_1^{(k)} - 2\, w_2^{(k)}}{5} \end{array} \quad = \begin{bmatrix} 0 & 0 & -2/3 \\ 0 & 0 & -2/3 \\ -2/5 & -2/5 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^{(k)} + \begin{bmatrix} (\sqrt{3} - 4)/3 \\ -8/3 \\ -2/5 \end{bmatrix} .$$

Donde, $w^{(1)} = ((\sqrt{3} - 8)/3, -4, -4/5)^T$.
Assim,

$$w^{(1)} - w^{(0)} = ((\sqrt{3} - 11)/3, -4, -14/5) \Longrightarrow ||v^{(1)} - v^{(0)}||_\infty = 4 .$$

Atendendo a que $||C_J||_\infty = max(2/3, 2/3, 4/5) = 4/5$, resulta

$$||w - w^{(2)}||_\infty \leq \frac{||C_J||_\infty^2}{1 - ||C_J||_\infty} ||w^{(1)} - w^{(0)}||_\infty = 5 \times (4/5)^2 \times 4 .$$

3(c) Para

$$B = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 2 \\ 2 & 2 & 5 \end{bmatrix},$$

tem-se,

$$||B||_\infty = ||B||_1 = max(5, 5, 9) = 9 .$$

Assim,

$$cond_\infty(B) = ||B||_\infty \, ||B^{-1}||_\infty = 9 \times (9/9) = 9 .$$

O número de condição dá indicação da maior ou menor sensibilidade do sistema $B\, w = b$ a perturbações nos dados. Um número de condição muito superior a 1 indica que o sistema pode ser muito sensível a erros (por exemplo de arredondamento) quer no segundo membro, ou na matriz, ou em ambos.

4) Sendo

$$f(x_1, x_2, x_3) = (3\, x_1 + x_3^2 - \sqrt{3}, 3\, (x_2 + 1) + x_3^2 - 11, 2\, x_1 + x_2\, (x_3 + 1) - 10),$$

tem-se, para $x^{(0)} = (1, 5, 1)^T$,

$$J_f(x_1, x_2, x_3) = \begin{bmatrix} 3 & 0 & 2\, x_3 \\ 0 & 3 & 2\, x_3 \\ 2 & x_3 + 1 & x_2 \end{bmatrix}_{|x^{(0)}} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 2 \\ 2 & 2 & 5 \end{bmatrix} = B .$$

Como $J_f(x^{(0)})\Delta x^{(0)} = -f(x^{(0)})$, onde $f(x^{(0)}) = (4 - \sqrt{3}, 8, 2)^T$, o sistema linear a resolver é $B\, w = -f(x^{(0)}) = b$. Atendendo a que $w^{(1)} = ((\sqrt{3} - 8)/3, -4, -4/5)^T$, resulta

$$x^{(1)} = x^{(0)} + \Delta x^{(0)} \simeq x^{(0)} + w^{(1)} = ((\sqrt{3} - 5)/3, 1, 1/5)^T .$$

A.2.19

Exame de 12 de Janeiro de 2015 (Parte 1)

1) Considere um sistema de ponto flutuante $FP(10, 4, -10, 10)$, com arredondamento por corte. É dado o número real $x = 314.15162 \times 10^{-2}$ (próximo de π).

[1.0] **(a)** Ao calcular-se o valor $y = \sin(x)$, no referido sistema, observou-se um grande erro relativo para o resultado. Compare o erro relativo do valor arredondado, $fl(x)$, com o erro relativo aproximado do valor calculado para y (dando esses erros expressos em percentagem).

[1.0] **(b)** Dê uma explicação para a observação referida na alínea anterior, recorrendo ao número de condição da função em causa.

2) Considere a equação $f(x) = 0$, a qual possui um única raiz $z \in [0, 3/2]$, sendo

$$f(x) = x^3 + 2\,x^2 + 9\,x - 15 \ .$$

[1.5] **(a)** Se escolher para aproximações iniciais os valores $x_1 = 3/2$ e $x_2 = 1/2$, poderá garantir convergência do método da secante para a raiz z considerada? Justifique.

[1.5] **(b)** Calcule duas iteradas do método referido na alínea anterior e majore os respectivos erros absolutos.

[1.5] **(c)** Considere o método iterativo $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, com $x_0 = 1$, gerado por uma função da forma $g(x) = x - f(x)/f'(1)$. Mostre que a sucessão $(x_k)_{k>0}$ é convergente para z.

[1.0] **(d)** Para o processo iterativo considerado em **(c)** obtenha uma aproximação da respectiva constante assimptótica de convergência. Poderá afirmar que a sucessão em causa possui convergência linear? Justifique.

3) Sendo $a \neq 0$ e $b \in \mathbb{R}$ parâmetros reais, considere o sistema linear

$$\begin{cases} a\,x_1 + b\,x_2 + x_3 &= 0 \\ 3\,x_1 + x_2 &= 0 \\ b\,x_2 + a\,x_3 &= 0 \ . \end{cases}$$

[1.0] **(a)** Admitindo que o sistema possui solução única, escolha os parâmetros a e b de modo a garantir convergência do método de Jacobi para tal solução. Justifique a escolha que fizer.

[1.5] **(b)** Faça $a = 1$ e $b \in \mathbb{R}$ (qualquer). Depois de verificar que o sistema possui uma só solução, mostre que o método de Gauss-Seidel produz a solução em duas iterações, independentemente da escolha da aproximação inicial $x^{(0)}$ que considerar, caso sejam efectuados cálculos exactos. Justifique.

Resolução

1(a) O número x dado é representado no sistema de ponto flutuante por $\bar{x} = fl(x) = +0.3141 \times 10^1$. Logo,

$$|\delta_{\bar{x}}| = \frac{|x - \bar{x}|}{|x|} = \frac{0.0005162}{3.1415162} \simeq 0.000164 = 0.0164\,\% \ .$$

Mas,

$$\begin{aligned} \bar{y} = \sin(\bar{x}) &\simeq 0.00059265356 \\ y = \sin(x) &\simeq 0.00007645359, \end{aligned}$$

$$|\delta_{\bar{y}}| = \frac{|y - \bar{y}|}{|y|} \simeq \frac{0.0005162}{0.00007645} \simeq 6.8 = 680\,\% \ .$$

1(b) Na alínea anterior a um pequeno erro relativo $|\delta_{\bar{x}}|$ corresponde um grande erro relativo no resultado $\bar{y}$. Tal fica a dever-se ao mau condicionamento da função $\sin(x)$ para valores do argumento próximos de π. Com efeito,

$$cond_f(x) = \left| \frac{x\, f'(x)}{f(x)} \right| = \left| \frac{x\, \cos(x)}{\sin(x)} \right| ,$$

donde

$$\lim_{x \to \pi} cond_f(x) = +\infty \ .$$

Assim, $|\delta_{\bar{y}}| \simeq cond_f(\bar{x})\, |\delta_{\bar{x}}| > 1$, confirmando-se que o valor $\bar{y}$ calculado na alínea anterior está necessariamente muito contaminado pelo erro de arredondamento propagado pela função, o qual é muito ampliado.

2(a) Seja $I = [0, 3/2]$. A função e as suas derivadas são contínuas em I. Tem-se

$$f(0) = -15 < 0, \quad f(3/2) = 51/8 = 6.375 > 0,$$

e

$$\begin{aligned} f'(x) &= 3\,x^2 + 4\,x + 9 \neq 0 \quad \forall x \in I \\ f''(x) &= 6\,x + 4 > 0 \quad \forall x \in I \ . \end{aligned}$$

Note-se que a equação $f'(x) = 0$ não possui raízes reais. Como $f'' > 0$, conclui-se que f' é função estritamente crescente no intervalo. Uma vez que $f'(0) = 9 > 0$, esta função é positiva e monótona em I. Sabemos que $f \in C^2(I)$, $f(0) \times f(3/2) < 0$, e ambas as funções f' e f'' mantém sinal (positivo) no intervalo em causa. Considerando o subintervalo $[x_2, x_1] = [1/2, 3/2]$, tem-se

$$\left| \frac{f(1/2)}{f'(1/2)} \right| = \left| \frac{-79}{94} \right| \simeq 0.84 < 1 \quad \text{e} \quad \left| \frac{f(3/2)}{f'(3/2)} \right| = \left| \frac{17}{58} \right| \simeq 0.29 < 1,$$

podemos concluir que o método da secante é convergente para z, uma vez escolhidos x_1 e $x_2 \in I$.

2(b) Para $x_2 = 1/2$, $f(x_2) = -79/8 < 0$, logo $z \in (x_2, x_1) \Longrightarrow |z - x_2| < |x_2 - x_1| = 1$.

$$x_3 = x_2 - f(x_2) \frac{x_2 - x_1}{f(x_2) - f(x_1)} = \frac{72}{65} \simeq 1.1076923 \ .$$

Dado que $f(x_3) \simeq -1.22 < 0$, tem-se

$$z \in (x_3, x_1) \Longrightarrow |z - x_3| < |x_3 - x_1| \simeq 0.39230769 \qquad (*)$$

$$x_4 = x_3 - f(x_3)\frac{x_3 - x_2}{f(x_3) - f(x_2)} \simeq 1.1931667 \ .$$

Visto que $f(x_4) \simeq 0.28 > 0$, $z \in (x_3, x_4) \Longrightarrow |z - x_4| < |x_4 - x_3| \simeq 0.0855$.

Uma majoração de erro mais grosseira poderá ser obtida do seguinte modo. Seja

$$M = \frac{1}{2}\,\frac{max\,|f^{(2)}(x)|}{min_{0\le x\le 3/2}|f'(x)|} = \frac{f^{(2)}(3/2)}{2\,f'(0)} = \frac{1}{2} \times \frac{13}{9} \simeq 0.72 \ .$$

Dado que

$$z - x_4 = -\frac{1}{2}\,\frac{f^{(2)}(\xi_4)}{f'(\eta_4)}\,(z - x_3)\,(z - x_2), \quad \xi_4, \eta_4 \in (0, 3/2),$$

resulta

$$|z - x_4| \le M\,|z - x_3|\,|z - x_2|, \quad \text{onde} \quad x_2 = 1/2 \ .$$

Uma vez que $f(x_2) < 0$ e $f(3/2) > 0$, a raiz pertence ao subintervalo $(x_2, x_1) = (1/2, 3/2)$. Por conseguinte, $|z - x_2| < 1$. Consequentemente, atendendo a $(*)$, obtém-se

$$|z - x_4| < M \times 0.39230769 \times 1 \simeq 0.282 \ .$$

2(c) A função iteradora

$$g(x) = x - \frac{f(x)}{16} = \frac{16\,x - f(x)}{16} \quad \in C^\infty(\mathbb{R}) \ .$$

Dado que

$$f(x_0) = f(1) = -3 < 0 \quad \text{e} \quad f(3/2) = 51/8 > 0,$$

o (único) zero de f localiza-se no intervalo $I = [1, 3/2]$. Como $g(z) = z$, a raiz z é ponto fixo da função iteradora g . De

$$g'(x) = \frac{16 - f'(x)}{16} \quad \Rightarrow \quad g''(x) = \frac{-(6\,x + 4)}{16} < 0 \quad \forall x \in I,$$

resulta que g' é estritamente decrescente em I, com $g'(1) = 0$ e satisfazendo as desigualdades

$$-0.36 \simeq -23/64 = g'(3/2) \le g'(x) < 0, \quad x \in I \qquad (**)$$

Atendendo a que $g(1) = 19/16$ e $g(3/2) = 141/128$ e g é estritamente decrescente em I, resulta

$$1 < g(3/2) \le g(x) \le g(1) < 3/2 \ .$$

Conclui-se que $g(I) \subset I$ e $max_{x\in I}|g'(x)| = 23/64 < 1$. O teorema do ponto fixo é válido, pelo que escolhido $x_0 = 1 \in I$, a sucessão de iteradas converge para o ponto fixo z .

2(d) As desigualdades $(**)$ na alínea anterior indicam que $0 < |g'(x)| < 1$, pelo que a convergência do método é linear. Tem-se,

$$\lim_{k\to\infty} \frac{|z - x_{k+1}|}{|z - x_k|} = |g'(z)| \neq 0 \ .$$

Tomando como aproximação de z, por exemplo, o valor calculado na alínea **(b)**, $z \simeq x_4 \simeq 1.193$, obtém-se

$$g'(z) \simeq \frac{16 - f'(1.193)}{16} \simeq -0.128 \neq 0 \ .$$

3(a) A matriz de iteração do método é

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -b/a & -1/a \\ -3 & 0 & 0 \\ 0 & -b/a & 0 \end{bmatrix} \ .$$

A respectiva equação característica $det(C_J - \lambda I) = 0$, escreve-se

$$\lambda^3 - 3(b/a\,\lambda - b/a^2) = 0 \ .$$

Por exemplo, para $b = 0$ e $a = 1$, resulta $\lambda^3 = 0$, caso em que o raio espectral de C_J é nulo. Como para esta escolha de parâmetros a solução (única) é $x_1 = x_2 = x_3 = 0$, podemos concluir que o método converge para a solução do sistema, independentemente da escolha da aproximação inicial.

3(b) O sistema pode escrever-se na forma $A\,x = b$, onde

$$A = \begin{bmatrix} 1 & b & 1 \\ 3 & 1 & 0 \\ 0 & b & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \ .$$

Notar que $det(A) = -1 \neq 0$, pelo que a solução (única) do sistema é $x = (0,0,0)^T$. As fórmulas computacionais do método, escrevem-se

$$\begin{cases} x_1^{(k+1)} = -b\,x_2^{(k)} - x_3^{(k)} \\ x_2^{(k+1)} = -3\,x_1^{(k+1)} = 3\,b\,x_2^{(k)} + 3\,x_3^{(k)}, \qquad\qquad k = 0,1,\ldots \\ x_3^{(k+1)} = -b\,x_2^{(k+1)} = -3\,b^2\,x_2^{(k)} - 3\,b\,x_3^{(k)} \ . \end{cases}$$

Assim, o método é da forma $x^{(k+1)} = C_{GS}\,x^{(k)}, \quad k = 0,1,\ldots$, onde

$$C_{GS} = \begin{bmatrix} 0 & -b & -1 \\ 0 & 3\,b & 3 \\ 0 & -3\,b^2 & -3\,b \end{bmatrix}.$$

Seja $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^T$ uma qualquer aproximação inicial da solução. Tem-se,

$$\begin{aligned} x^{(1)} &= C\,x^{(0)} = (-b\,x_2^{(0)} - x_3^{(0)}, 3\,b\,x_2^{(0)} + 3\,x_3^{(0)}, -3\,b^2\,x_2^{(0)} - 3\,b\,x_3^{(0)})^T \\ x^{(2)} &= C\,x^{(1)} = (0,0,0)^T = x \ . \end{aligned}$$

A.2.20

Exame de 12 de Janeiro de 2015 (Parte 2)

1. Considere a seguinte tabela de valores de uma função f, de classe C^3 em $I = [1, 8]$,

x_i	1	2	4	7	8
$f(x_i)$	1.2	1.5	2	1.4	0.5

Seja z um número inteiro, tal que $z \neq 2$ e $f(z) = 1.5$.

[1.0] **(a)** Mostre que $f[1, z, 2] = \dfrac{0.3}{1 - z}$, onde $f[1, z, 2]$ designa uma diferença dividida de segunda ordem.

[1.0] **(b)** Supondo que, para $x \leq 2$, a função f tem a forma

$$f(x) = -x^2 + a\,x + b,$$

determine z, atendendo à igualdade da alínea anterior.

[1.0] **(c)** Através da fórmula de Newton com diferenças divididas, construa o polinómio de grau ≤ 2 que lhe permite obter a melhor aproximação para o valor da função em $x = 6$. Justifique.

[1.5] **(d)** Calcule um valor aproximado de $f(6)$ através do polinómio referido na alínea **(c)**, e obtenha uma estimativa para o erro absoluto que comete nessa aproximação.

[1.5] **2.** Sendo a, b e c parâmetros reais, utilize o método dos mínimos quadrados para ajustar uma função da forma $g(x) = \dfrac{1}{a + b\,x} + c$, à seguinte tabela de valores de uma função f, sabendo-se que $\lim_{x\to\infty} f(x) = 20$:

x	0	2	4	6	8	10
$f(x)$	84.8	75.0	67.2	61.9	57.6	53.4

(Indique os valores que calcular para as entradas do respectivo sistema de equações normais; não é necessário resolver o sistema).

3. Considere o integral $\displaystyle\int_0^1 \frac{1}{1 + 2\,x}\,dx$.

[1.0] **(a)** Recorrendo à regra dos trapézios composta, determine o número mínimo de subintervalos necessários para garantir um erro absoluto inferior a 10^{-10}.

[1.0] **(b)** O mesmo que a alínea anterior para a regra de Simpson composta.

4. Considere o problema de valores iniciais

$$\begin{cases} y'' = y + e^t, & t \in [0, 0.2] \\ y(0) = 1 \\ y'(0) = 0\,. \end{cases}$$

[0.5] **(a)** Reduza-o a um sistema de equações de primeira ordem.

[1.5] **(b)** Para o passo $h = 0.1$, obtenha um valor aproximado de $y'(0.2)$, usando o método de Euler explícito.

Resolução

1(a)

$$f[1, z, 2] = \frac{f[z, 2] - f[1, z]}{2 - 1} = f[z, 2] - f[z, 1] = \frac{1.5 - 1.5}{2 - z} - \frac{1.5 - 1.2}{z - 1} = \frac{0.3}{1 - z} \, .$$

1(b) Dado que para $x \leq 2$ se tem $f(x) = p_2(x)$, resulta

$$f(x) = p_2(x) = f(1) + f[1, z]\,(x - 1) + f[1, z, 2])\,(x - 1)(x - 2) \, .$$

Assim, o coeficiente do termo de maior grau (para a função f e para o polinómio interpolador), satisfaz a relação $f[1, z, 2] = -1$. Por conseguinte,

$$\frac{0.3}{1 - z} = -1 \quad \Leftrightarrow \quad 0.3 = z - 1 \quad \Leftrightarrow \quad z = 1.3 \, .$$

1(c) A fim de minimizar o erro de interpolação, considerem-se os 3 pontos tabelados mais próximos de $x = 6$, ou seja, $x_0 = 4$, $x_1 = 7$ e $x_2 = 8$. A respectiva tabela de diferenças divididas é:

x	$f(x)$	$f[.,.]$	$f[.,.,.]$
4	2		
		-0.2	
7	1.4		
			-0.175
		-0.9	
8	0.5		

Donde,

$$p_2(x) = 2 - 0.2\,(x - 4) - 0.175\,(x - 4)\,(x - 7) \, .$$

1(d)

$$f(6) \simeq p_2(6) = 2 - 0.2 \times 2 - 0.175 \times 2 \times (-1) = 1.95 \, .$$

Atendendo à fórmula de erro de interpolação,

$$\begin{aligned} |e_2(6)| &= |f(6) - p_2(6)| = \left| \frac{f^{(3)}(\xi)}{3!}\,(6 - x_0)\,(6 - x_1)\,(6 - x_2) \right| \\ &= 4 \left| \frac{f^{(3)}(\xi)}{3!} \right|, \qquad \xi \in (4, 8) \, . \end{aligned}$$

O valor $\dfrac{f^{(3)}(\xi)}{3!}$ pode ser estimado através da diferença dividida de terceira ordem $f[4,7,8,x] = f[4,6,8,2]$,

x	$f(x)$	$f[.,.]$	$f[.,.,.]$	$f[.,.,.,.]$
4	2			
		-0.2		
7	1.4			
			-0.175	
		-0.9		$-0.0141(6)$
8	0.5		$-0.146(6)$	
		$-0.166(6)$		
2	1.5			

Ou seja,

$$\left|\frac{f^{(3)}(\xi)}{3!}\right| \simeq |f[4,7,8,2]| \simeq 0.0142 \ .$$

Por conseguinte,

$$|e_2(6)| \simeq 4 \times 0.0142 \simeq 0.06 \ .$$

2. Atendendo a que $\lim_{x\to\infty} \dfrac{1}{a+b\,x} + c = c$, admitimos que

$$\lim_{x\to\infty} g(x) = \lim_{x\to\infty} f(x) = c = 20 \ .$$

Assim,

$$g(x) \simeq \frac{1}{a+b\,x} + 20 \quad \Longrightarrow g(x) - 20 \simeq \frac{1}{a+b\,x} \ .$$

Seja

$$G(x) = \frac{1}{g(x)-20} \simeq a + b\,x = a\,\phi_0(x) + b\,\phi_1(x)\,.$$

Para $F = (1/(f(x_0)-20), \ldots, 1/(f(x_5)-20))^T \simeq (0.0154, 0.0182, 0.0212, 0.0239, 0.0266, 0.0299)^T$, obtém-se o sistema de equações normais

$$\begin{bmatrix} 6 & 30 \\ 30 & 220 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.135 \\ 0.776 \end{bmatrix} \ .$$

3(a) Atendendo a que o comprimento do intervalo vale $b-a=1$ e

$$f^{(2)}(x) = \frac{8}{(1+2\,x)^3} \quad \Longrightarrow \quad max_{x\in[0,1]}|f^{(2)}(x)| = f^{(2)}(0) = 8,$$

tem-se

$$|E_N^T(f)| \leq \frac{8\,h^2}{12} < 10^{-10} \quad \Longleftrightarrow \quad h < 10^{-5}\sqrt{3/2} \simeq 0.000012247\,.$$

Ou seja,

$$N > 1/h \simeq 81649.7 \quad \Longrightarrow N = 81650 \ .$$

3(b)

$$f^{(4)}(x) = \frac{384}{(1+2\,x)^5} \quad \Longrightarrow \quad max_{x\in[0,1]}|f^{(4)}(x)| = f^{(4)}(0) = 384,$$

$$|E_N^S(f)| \leq \frac{384\,h^4}{180} < 10^{-10} \quad \Longleftrightarrow \quad h < (180\times 10^{-10}/384)^{1/4} \simeq 0.0026165878\ .$$

Assim, dado que $N > 1/h \simeq 382.2$ deverá ser número natural par, deverão considerar-se pelo menos 384 subintervalos.

4(a) Fazendo $y_1 = y$ e $y_2 = y'$, resulta o sistema

$$\left\{\begin{array}{ll} y_1' & = y_2, \\ y_2' & = y_1 + e^t, \end{array}\right. \qquad t \in [0, 0.2]$$

de valores iniciais $y_{1,0} = y(0) = 1$ e $y_{2,0} = 0$.

4(b) Para o método de Euler explícito aplicado ao sistema da alínea anterior, resulta

$$\left\{\begin{array}{ll} y_{1,i+1} & = y_{1,i} + h\,y_{2,i}, \\ y_{2,i+1} & = y_{2,i} + h\,(y_{1,i} + e^{t_i}), \end{array}\right. \qquad i = 0, 1.$$

com $y_{1,0} = 1$ e $y_{2,0} = 0$. Para aproximar $y'(0.2)$, efectuam-se os dois passos indicados a seguir.
Para $t_0 = 0$,

$$\left\{\begin{array}{ll} y_{1,1} & = y_{1,0} + h\,y_{2,0} = 1 + 0.1\times 0 = 1 \simeq y(0.1) \\ y_{2,1} & = y_{2,0} + h\,(y_{1,0} + e^{t_0}) = 0 + 0.1\,(1+1) = 0.2 \simeq y'(0.1), \end{array}\right.$$

$t_1 = t_0 + h = 0.1$,

$$\left\{\begin{array}{ll} y_{1,2} & = y_{1,1} + h\,y_{2,1} = 1 + 0.1\times 0.2 = 1.02 \simeq y(0.2) \\ y_{2,2} & = y_{2,1} + h\,(y_{1,1} + e^{t_1}) = 0.2 + 0.1\,(1 + e^{0.1}) = 0.41051709 \simeq y'(0.2)\ . \end{array}\right.$$

A.2.21

(Teste de 8 de Abril de 2015)

1) Num sistema decimal de ponto flutuante com 6 dígitos na mantissa, considere os números $x = 121$ e $y = 1201$. Representando por $\tilde{x}$ e $\tilde{y}$, o número do sistema mais próximo e superior respectivamente a x e y, diga se é verdade que $\tilde{x} - x > \tilde{y} - y$. Justifique. **[1.0]**

2) Considere a função real

$$f(x) = 2\,x - |\,\cos(x)\,|\ .$$

(a) Mostre que a equação $f(x) = 0$ possui uma só raiz $\alpha \in (0, \pi/4)$. Calcule α, com erro absoluto inferior a 0.25, aplicando o método da bissecção. Justifique. **[1.5]**

(b) Prove que o processo iterativo **[1.0]**

$$x_{n+1} = \frac{\cos(x_n)}{2}, \qquad n = 0, 1, \ldots$$

converge para α independentemente da escolha que fizer de $x_0 \in (0, \pi/4)$.

[1.5] **(c)** Obtenha uma estimativa da constante assimptótica de convergência.
A convergência é monótona? Justifique.

[1.5] **(d)** Recorrendo ao método da secante, poderá garantir convergência local deste método para a raiz α em causa? Justifique.

[1.0] **3)** Qual a ordem de convergência do método de Newton quando aplicado para aproximar a raiz real da equação $2\,x^3 = 1$? Justifique.

4) Considere a matriz de entradas reais,

$$H = \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix} .$$

[1.0] **(a)** Para $c = -1/2$, obtenha $||H||_2$.

[1.5] **(b)** Para $c \neq 1$, calcule $cond_\infty(H)$, isto é, o número de condição da matriz H na norma ∞ . O que pode dizer sobre o condicionamento de um sistema da forma $H\,x = v$, quando c se aproxima de 1 ?

Resolução

1)

$$\begin{aligned} x &= 0.121000 \times 10^3 \\ \tilde{x} &= 0.121001 \times 10^3 \implies \tilde{x} - x = 10^{-3} \end{aligned}$$

$$\begin{aligned} y &= 0.120100 \times 10^4 \\ \tilde{x} &= 0.120101 \times 10^4 \implies \tilde{y} - y = 10^{-2} \ . \end{aligned}$$

Por conseguinte, a desigualdade do enunciado é falsa.

2(a) Seja $I = [0, \pi/4]$. Dado que $f(x) = 2\,x - \cos(x)$, para $0 \leq x \leq \pi/4$, é continua e diferenciável em I, sendo

$f(0) = -1 < 0$
$f(\pi/4) = 1/2 > 0 \implies$ existe pelo menos um zero de f em I;
$f'(x) = 2 + \sin(x) > 0 \ \forall x \in I \implies$ existe no máximo um zero de f em I .

Logo, existe exactamente um zero α de f no intervalo considerado. Aplicando o método da bissecção, resulta

$$\begin{aligned} &x_1 = \pi/8 \quad \text{e} \quad f(x_1) \simeq -0.14 \implies \alpha \in (\pi/8, \pi/4) \\ &|\alpha - x_1| < \pi/8 \simeq 0.39 > 0.25 \end{aligned}$$

$$\begin{aligned} &x_2 = \ = \frac{3\,\pi}{16} \simeq 0.58905 \quad \text{e} \\ &|\alpha - x_2| < \pi/16 \simeq 0.196 < 0.25 \ . \end{aligned}$$

2(b) A função iteradora $g(x) = \cos(x)/2 \in C^1([0, \pi/4])$. Tem-se,

$$g'(x) = -\frac{\sin(x)}{2} \leq 0 \quad \forall x \in I,$$
$$g(0) = 1/2 \quad \text{e} \quad g(\pi/4) = 1/4 \ .$$

Como a função g é positiva e decrescente em I, resulta

$$0 < 1/4 = g(\pi/4) \leq g(x) \leq g(0) = 1/2 < \pi/4 \quad \Leftrightarrow \quad g(I) \subset I \ .$$

Além disso,

$$|g'(x)| \leq 1/2 = L < 1, \quad \forall x \in I \ .$$

Pelo teorema do ponto fixo existe um só valor $z \in I$ tal que $g(z) = z$, e o método iterativo gerado por z converge (globalmente) para z. Assim, $2\,z = \cos(z) \Longrightarrow 2\,z - cos(z) = f(z) = 0$, isto é, z é zero de f em I, pelo que $z = \alpha$.

2(c) Sabe-se que

$$\lim_{k\to\infty} \left| \frac{\alpha - x_{k+1}}{\alpha - x_k} \right| = |g'(z)| = \frac{\sin(\alpha)}{2} \neq 0 \ .$$

Visto que $\alpha \simeq x_2$ (aproximação de α calculada em **2 (a)**), uma aproximação da constante assimpótica de erro é

$$k_\infty \simeq |g'(x_2)| = \sin(x_2)/2 \simeq 0.28 \ .$$

Atendendo a que $-1 < g'(x) \leq 0, \quad \forall x \in I$, a convergência do método é alternada.

2(d) Sim. Porquanto, uma vez que $f \in C^2(I)$, sendo α zero simples de f, sabe-se que uma vez escolhidos x_{-1} e x_0 suficientemente próximos de α, se tem para a sucessão de iteradas (x_k) do método da secante,

$$\lim_{k\to\infty} \left| \frac{\alpha - x_{k+1}}{(\alpha - x_k)^p} \right| = k_\infty > 0$$

onde $p = (1 + \sqrt{5})/2$, ou seja, a convergência é supralinear.

3) A equação considerada tem raiz única $z = 1/2^{1/3}$. Para $f(x) = 2\,x^3 - 1$, resulta $f'(z) \neq 0$. Assim, z é zero simples e o método converge quadraticamente, uma vez escolhida uma aproximação inicial x_0 suficientemente próxima de z .

4)(a) A matriz H é simétrica pelo que $||H||_2 = \rho(H)$. Ora,

$$det(H - \lambda I) = 0 \quad \Leftrightarrow \lambda = 1 \pm 1/2 \ .$$

Logo, $\rho(A) = 3/2$.

4)(b)

$$H^{-1} = \frac{1}{1 - c^2} \begin{bmatrix} 1 & -c \\ -c & 1 \end{bmatrix} \ .$$

Assim, $||H||_\infty = 1 + |c|$ e $||H^{-1}||_\infty = \dfrac{1 + |c|}{|1 - c^2|}$. Por conseguinte,

$$\lim_{c\to 1} cond_\infty(H) = \lim_{c\to 1} \frac{(1 + |c|)^2}{|1 - c^2|} = +\infty \ .$$

O número de condição poderá tomar valores muito elevados quando c for próximo de 1, o que significa que nesse caso a matriz é mal condicionada.

A.2.22

Teste de 28 de Maio de 2015

1) Considere os seguintes sistemas lineares (S_1) e (S_2),

$$(S_1) \rightarrow \quad A\,x = b \qquad \text{e} \qquad (S_2) \rightarrow \quad \frac{(A + A^T)}{2}\,u = b,$$

sendo

$$A = \begin{bmatrix} 4 & 0 & 0 \\ -2 & 4 & 0 \\ 0 & -2 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} \delta \\ 0 \\ 0 \end{bmatrix},$$

onde A^T designa a matriz transposta de A e δ é um número real.

[1.0] **(a)** Para o sistema (S_1), com $\delta \in \mathbb{R}$, escreva as fórmulas computacionais do método de Gauss-Seidel. Mostre que (no caso de não ocorrerem erros de arredondamento) $||x - x^{(1)}||_1 = 0$, independentemente da aproximação inicial $x^{(0)}$ que considerar.

[1.0] **(b)** Sendo $\delta \in \mathbb{R}$, se partir da aproximação inicial $u^{(0)} = (-10, 0, -10)^T$, o método de Jacobi aplicado ao sistema (S_2) é convergente? Justifique.

[1.0] **(c)** Para $\delta = 1$ e $u^{(0)} = (-1, 0, -1)^T$, obtenha um majorante de $||u - u^{(2)}||_\infty$, onde $u^{(2)}$ designa a segunda iterada do método de Jacobi aplicado ao sistema (S_2) .

[1.5] **2)** Dados os nós de interpolação $x_0 = -1$, $x_1 = 0$ e $x_2 = 1$, e uma tolerância $\epsilon > 0$, considere o polinómio $\bar{p}(x)$ interpolador de Lagrange da tabela $\{x_i, \bar{f}_i\}$, sendo $\bar{f}_0 = \bar{f}(x_0)$, $\bar{f}_1 = \bar{f}(x_1)$ e $\bar{f}_2 = \bar{f}(x_2)$ valores aproximados de uma função contínua f, tal que

$$max_{i=0,1,2}\ |f(x_i) - \bar{f}_i| \leq \epsilon\ .$$

Obtenha um majorante do erro $|p(1/2) - \bar{p}(1/2)|$, onde $p(x)$ designa o polinómio interpolador construido a partir dos valores exactos da função. Apresente todos os cálculos que efectuar.

3) Considere uma tabela de valores $\tau = \{x_i, \Psi(x_i)\}$, tal que $x_i = 1 + i\,h$, para $i = 0, 1, 2, 3$, onde $h = 0.2$ e $\Psi(x) = 1/\cos(x)$.

[1.5] **(a)** Sendo $N\,u = f$ o sistema de equações normais a partir do qual se pode calcular a melhor aproximação polinomial quadrática da tabela τ (no sentido dos mínimos quadrados), obtenha a matriz N efectuando cálculos exactos. Poderá usar o método de Gauss-Seidel para aproximar a solução de tal sistema? Justifique.

[1.5] **(b)** Usando os dois primeiros nós da tabela τ, aplique o método dos coeficientes indeterminados para obter uma aproximação do integral $\int_1^{1.2} \cos^{-1}(x)\,dx$. Qual é a designação habitual da regra de quadratura que construiu? Justifique.

4) Sejam $y' = 4\,e^{0.8\,x} - 0.5\,y$ e $y(0) = 2$, com $0 \leq x \leq 4$.

(a) Diga, justificando, se o problema de valor inicial dado possui solução única. **[1.0]**

(b) Aplique o método de Heun ($y_{i+1} = y_i + h/2\,[f(t_i, y_i) + f(t_i + h, y_i + h\,f(t_i, y_i))]$), com passo $h = 1$, para aproximar $y(2)$. **[1.5]**

Resolução

1 (a) As fórmulas computacionais do método escrevem-se

$$\left\{ \begin{array}{ll} x_1^{(k+1)} & = \delta/4 \\ x_2^{(k+1)} & = \dfrac{2\,x_1^{(k+1)}}{4} = \dfrac{\delta}{8}, \\ \\ x_3^{(k+1)} & = \dfrac{2\,x_2^{(k+1)}}{4} = \dfrac{\delta}{16}\ . \end{array} \right. \qquad k = 0, 1, \ldots$$

Assim, para qualquer aproximação inicial $x^{(0)}$, resulta $x^{(1)} = (\delta/4, \delta/8, \delta/16)^T$.
Uma vez que o sistema $A\,x = b$ é triangular, a respectiva solução obtém-se imediatamente: $x = (\delta/4, \delta/8, \delta/16)^T$. Por conseguinte, $||x - x^{(1)}||_1 = 0$, isto é, caso não haja lugar a erros de arredondamento, a solução exacta é alcançada após uma iteração do método de Gauss–Seidel.

1 (b) O sistema (S_2) é tridiagonal simétrico:

$$\left\{ \begin{array}{ll} 4\,u_1 - u_2 & = \delta = 1 \\ -u_1 + 4\,u_2 - u_3 & = 0 \\ -u_2 + 4\,u_3 & = 0\ . \end{array} \right.$$

A respectiva matriz de iteração do método de Jacobi escreve-se,

$$C_J = \left[\begin{array}{ccc} 0 & 1/4 & 0 \\ 1/4 & 0 & 1/4 \\ 0 & 1/4 & 0 \end{array} \right]\ .$$

Logo, $||C_J||_\infty = ||C_J||_1 = 1/2 < 1$. Por conseguinte, o método é convergente para a solução u, independentemente da aproximação inicial (em particular para a aproximação $u^{(0)}$ dada).

1 (c) A partir do sistema considerado na alínea anterior resultam de imediato as fórmulas computacionais

$$\left\{ \begin{array}{ll} u_1^{(k+1)} & = \dfrac{\delta + u_2^{(k)}}{4} = \dfrac{1 + u_2^{(k)}}{4} \\ u_2^{(k+1)} & = \dfrac{u_1^{(k)} + u_3^{(k)}}{4} \\ \\ u_3^{(k+1)} & = \dfrac{u_2^{(k)}}{4} \end{array} \right. \qquad k = 0, 1, \ldots$$

Assim, para $u^{(0)} = (-1, 0, -1)^T$, obtém-se

$$u^{(1)} = (1/4, -1/2, 0)^T$$
$$u^{(2)} = (1/8, 1/16, -1/8)^T \Longrightarrow u^{(2)} - u^{(1)} = (-1/8, 9/16, -1/8)^T \Longrightarrow ||u^{(2)} - u^{(1)}||_\infty = 9/16 \ .$$

Sabemos que $||C_J||_\infty = 1/2$. Por conseguinte,

$$\begin{aligned} ||u - u^{(2)}||_\infty &\leq \frac{||C_J||_\infty}{1 - ||C_J||_\infty} \, ||u^{(2)} - u^{(1)}||_\infty \\ &\leq ||u^{(2)} - u^{(1)}||_\infty = 9/16 = 0.5625 \ . \end{aligned}$$

2) A base de Lagrange associada aos nós de interpolação é constituída pelos polinómios de grau dois,

$$\begin{aligned} l_0(x) &= \frac{x\,(x-1)}{(-1)\,(-2)} \Longrightarrow l_0(1/2) = -1/8 \\ l_1(x) &= \frac{(x+1)\,(x-1)}{(-1)} \Longrightarrow l_1(1/2) = -3/4 \\ l_2(x) &= \frac{(x+1)x}{(-1)\,2} \Longrightarrow l_2(1/2) = 3/8 \ . \end{aligned}$$

Dado que

$$\begin{aligned} \bar{p}(x) &= \bar{f}_0 \, l_0(x) + \bar{f}_1 \, l_1(x) + \bar{f}_2 \, l_2(x) \\ p(x) &= f_0 \, l_0(x) + f_1 \, l_1(x) + f_2 \, l_2(x), \end{aligned}$$

resulta

$$\begin{aligned} |p(x) - \bar{p}(x)| &\leq |f_0 - \bar{f}_0| \, |l_0(x)| + |f_1 - \bar{f}_1| \, |l_1(x)| + |f_2 - \bar{f}_2| \, |l_2(x)| \\ &\leq \epsilon \, (|l_0(1/2)| + |l_1(1/2)| + |l_2(1/2)|) \\ &\leq \frac{5}{4} \, \epsilon \ . \end{aligned}$$

3 (a) Trata-se de determinar a melhor aproximação de mínimos quadrados da tabela, por funções do tipo $g(x) = a_0 + a_1 \, x + a_2 \, x^2$. Fazendo $\phi_0 = (1, 1, 1, 1)^T$, $\phi_1 = (1, 1.2, 1.4, 1.6)^T$ e $\phi_2 = (1, 1.2^2, 1.4^2, 1.6^2)^T$, a melhor aproximação de mínimos quadrados resulta do vector $g = a_0 \phi_0 + a_1 \, \phi_1 + a_2 \, \phi_2$ cujos coeficientes são solução do sistema normal, de matriz simétrica,

$$N = \begin{bmatrix} < \phi_0, \phi_0 > & < \phi_0, \phi_1 > & < \phi_0, \phi_2 > \\ < \phi_0, \phi_1 > & < \phi_1, \phi_1 > & < \phi_1, \phi_2 > \\ < \phi_0, \phi_2 > & < \phi_1, \phi_2 > & < \phi_2, \phi_2 > \end{bmatrix} = \begin{bmatrix} 4 & 5.2 & 6.96 \\ 5.2 & 6.96 & 9.568 \\ 6.96 & 9.568 & 13.4688 \ . \end{bmatrix}$$

Dada que a matriz N é definida positiva, o método de Gauss-Seidel é convergente quando aplicado para aproximar a solução do sistema normal.

3 (b) Para $f(x) = cos^{-1}(x)$, $a = 1$, $b = 1.2$, pretende-se determinar a regra interpolatória de quadratura $Q(f) = A_0 \, f(a) + A_1 \, f(b)$, tal que $Q(1) = I(1) = \int_1^{1.2} dx$ e $Q(x) = I(x) = \int_1^{1.2} x \, dx$. O sistema a resolver é

$$\begin{bmatrix} 1 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} = \begin{bmatrix} I(1) \\ I(x) \end{bmatrix}.$$

Como $Q(f)$ é exacta para polinómios de grau ≤ 1 e usa como nós os extremos do intervalo, trata-se da regra dos trapézios. Assim, $A_0 = A_1 = (b-a)/2 = 0.1$. Por conseguinte, o integral $\int_1^{1.2} \cos^{-1}(x)\, dx$ é aproximado por $0.1\, \left(\cos^{-1}(1) + \cos^{-1}(1.2)\right) \simeq 0.46105$.

4 (a) Sejam $D = \{(x, y) : 0 \leq x \leq 4,\ y \in \mathbb{R}\}$ e a função f, definida em D,

$$f(x, y) = 4\, e^{0.8\, x} - 0.5\, y \ .$$

Como $f \in C^1(D)$ e $|\partial f(x, y)/\partial y| = 0.5 < \infty$, sabe-se que o problema de valor inicial dado tem solução única.

4 (b) Para $h = 1$, tem-se

$$\begin{aligned} f(x+1, y + f(x, y)) &= 4\, e^{0.8\,(x+1)} - 0.5\, (y + f(x, y)) \\ &= 4\, e^{0.8\,(x+1)} - 0.5\, (y + 4\, e^{0.8\, x} - 0.5\, y) \\ &= 4\, e^{0.8\,(x+1)} - 0.5\, (0.5\, y + 4\, e^{0.8\, x}) \ . \end{aligned}$$

Assim,

$$\begin{aligned} y_{i+1} &= y_i + 1/2\, \left[4\, e^{0.8\, x_i} - 0.5\, y_i + 4\, e^{0.8\,(x_i+1)} - 0.5\, (0.5\, y_i + 4\, e^{0.8\, x_i})\right] \\ &= y_i + 1/2\, \left[2\, e^{0.8\, x_i} + 4\, e^{0.8\,(x_i+1)} - 0.75\, y_i\right], \quad \text{para} \quad i = 0, 1 \ . \end{aligned}$$

Logo, como $x_0 = 0$ e $y_0 = 2$, resulta

$$\begin{aligned} y_1 &= 2 + 1/2\, \left[2 + 4\, e^{0.8} - 0.75 \times 2\right] \simeq 6.70108 \\ y_2 &= y_1 + 1/2\, \left[2\, e^{0.8} + 4\, e^{0.8 \times 2} - 0.75 \times y_1\right] \simeq 16.320 \simeq y(2) \ . \end{aligned}$$

A.2.23

Exame de 2 de Julho de 2015 (Parte II)

1) Seja $M\, x = c$ um sistema tridiagonal simétrico tal que $m_{i,i} = 4$, $m_{i,j} = -1$ se $|i - j| = 1$, (restantes entradas nulas) e $c_i = \sum_{j=1}^{4} m_{i,j}$, para $i = 1, \ldots, 4$. Como aproximação da solução do sistema tome $x^{(0)} = (-1, 0, 1, 0)^T$.

(a) Escreva as fórmulas computacionais do método de Gauss-Seidel aplicado ao sistema e diga, justificando, se o método é convergente caso inicie esse processo iterativo com o vector $x^{(0)}$. **[1.0]**

(b) Obtenha um majorante de $||x - x^{(1)}||_\infty$, onde x é solução de $M\, x = c$, e $x^{(1)}$ a primeira iterada do referido método, partindo da aproximação $x^{(0)}$ dada. **[1.0]**

(c) Admita que $A\, x = b$ é um sistema linear não singular, de três equações a três incógnitas, com entradas não nulas na diagonal principal. Considere o processo iterativo **[1.5]**

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{2} \left(b_i - \sum_{j=1}^{3} a_{i,j}\, x_j^{(k)} \right) / a_{i,i} \quad \text{para} \quad i = 1, 2, 3, \qquad k = 0, 1, \ldots \ .$$

Mostre que o processo é do tipo $x^{(k+1)} = C\,x^{(k)} + d$, calculando a matriz C e o vector d. Prove que no caso da matriz A ser de diagonal estritamente dominante por linhas, este método converge para a solução do sistema, independentemente da aproximação inicial $x^{(0)}$ que considerar.

2) Considere a tabela de valores $\Omega = \{x_i, \sin(x_i)\}$, onde $x_i = \dfrac{\pi}{2}\,(i-1)$, para $i = 0, 1, 2, 3$.

[1.0] **(a)** Efectuando cálculos exactos, obtenha a melhor aproximação de mínimos quadrados da tabela Ω mediante funções do tipo $g(x) = b\,x + c$, tal que $g(0) = 0$ e $b, c \in \mathbb{R}$. Justifique.

[1.5] **(b)** Designe por $q(x)$ o polinómio interpolador da tabela Ω, na forma de Lagrange. Mostre que existe $\theta \in (-\pi/2, \pi)$, para o qual $\dfrac{\sin(1/2) - q(1/2)}{K} = \sin(\theta)/24$, onde K é uma constante que deverá escrever exactamente.

3) Para $h > 0$ e f função diferenciável, pretende-se determinar $I(f) = \displaystyle\int_{-h/2}^{h/2} f(x)\,dx$.

[1.0] **(a)** Sendo $T(f)$ a regra dos trapézios simples, considere a regra de quadratura $R(f) = T(f) - \dfrac{h^2}{12}\left[f'(h/2) - f'(-h/2)\right]$. Mostre que a regra $R(f)$ é exacta para qualquer polinómio de grau não superior a três. Justifique.

[1.5] **(b)** Sendo $I(f) = \displaystyle\int_{-1/2}^{1/2} (e^x - 2)\,dx$, calcule o valor que se obtém aplicando a regra $R(f)$ e compare com o que resulta da regra de Newton–Cotes fechada, com três nós. Poderá dizer que o erro absoluto desta última regra é inferior ao da regra $R(f)$? Justifique.

4) Considere o problema de valor inicial $y'(x) = x\,\cos(x\,y(x))$, com $y(1) = -1$ e $1 \leq x \leq 3$.

[1.0] **(a)** Aplique o método de Euler modificado, $y_{i+1} = y_i + h\,f\,(t_i + h/2, y_i + h/2\,f(t_i, y_i))$, com passo $h = 1/2$, a fim de aproximar $y(2)$.

[0.5] **(b)** Suponha que para $1 \leq x \leq 2$ a solução do problema dado é $y(x) = -1 + 1/2\,(x - x^2)$. Nesse caso, ao aplicar o método anterior, qual o erro que poderá prever ao aproximar $y(2)$, com passo $h = 0.1$? Justifique.

Resolução

1 (a)

$$M\,x = c \iff \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 3 \end{bmatrix} .$$

Assim, as fórmulas computacionais do método escrevem-se:

$$\begin{aligned}
x_1^{(k+1)} &= \frac{3 + x_2^{(k)}}{4} \\
x_2^{(k+1)} &= \frac{2 + x_1^{(k+1)} + x_3^{(k)}}{4} = \frac{11 + x_2^{(k)} + 4\,x_3^{(k)}}{16} \\
x_3^{(k+1)} &= \frac{2 + x_2^{(k+1)} + x_4^{(k)}}{4} = \frac{43 + x_2^{(k)} + 4\,x_3^{(k)} + 16\,x_4^{(k)}}{64} \\
x_4^{(k+1)} &= \frac{3 + x_3^{(k+1)}}{4} = \frac{235 + x_2^{(k)} + 4\,x_3^{(k)} + 16\,x_4^{(k)}}{256} \ .
\end{aligned}$$

A matriz M é definida positiva (como de pode facilmente comprovar calculando os seus menores principais) e, por conseguinte, o método converge independentemente do vector inicial escolhido. Pode chegar-se à mesma conclusão atendendo a que das fórmulas anteriores resulta imediatamente o vector d bem como a seguinte matriz de iteração, com norma inferior a 1 :

$$d = (3/4, 11/16, 43/64, 235/256)^T$$

$$C_{GS} = \begin{bmatrix} 0 & 1/4 & 0 & 0 \\ 0 & 1/16 & 1/4 & 0 \\ 0 & 1/64 & 1/16 & 1/4 \\ 0 & 1/256 & 1/64 & 1/16 \end{bmatrix} \Longrightarrow ||C_{GS}||_\infty = 21/64 \simeq 0.328125 < 1 \ .$$

1 (b) Para $x^{(0)} = (-1, 0, 1, 0)^T$, resulta

$$\begin{aligned}
x^{(1)} &= (3/4, 15/16, 47/64, 239/256)^T = (0.75, 0.9375, 0.734375, 0.93359375)^T \\
x^{(1)} - x^{(0)} &= (1.75, 0.9375, -0.265625, 0.93359375)^T \ .
\end{aligned}$$

Logo,

$$||x^{(1)} - x^{(0)}||_\infty = 1.75$$

e

$$\begin{aligned}
||x - x^{(1)}||_\infty &\leq \frac{||C_{GS}||_\infty}{1 - ||C_{GS}||_\infty} \, ||x^{(1)} - x^{(0)}||_\infty \\
&\leq \frac{21}{43} \, ||x^{(1)} - x^{(0)}||_\infty \simeq 0.85 \ .
\end{aligned}$$

1 (c) O processo escreve-se

$$\left\{ \begin{array}{l}
x_1^{(k+1)} = x_1^{(k)} + b_1/(2\,a_{11}) - \left(\dfrac{a_{11}\,x_1^{(k)} + a_{12}\,x_2^{(k)} + a_{13}\,x_3^{(k)}}{2\,a_{11}} \right) \\
x_2^{(k+1)} = x_2^{(k)} + b_2/(2\,a_{22}) - \left(\dfrac{a_{21}\,x_1^{(k)} + a_{22}\,x_2^{(k)} + a_{23}\,x_3^{(k)}}{2\,a_{22}} \right) \\
x_3^{(k+1)} = x_3^{(k)} + b_3/(2\,a_{33}) - \left(\dfrac{a_{31}\,x_1^{(k)} + a_{32}\,x_2^{(k)} + a_{33}\,x_3^{(k)}}{2\,a_{33}} \right)
\end{array} \right. \ .$$

Assim, a respectiva matriz de iteração, seja C, é da forma

$$C = \begin{bmatrix} 1/2 & -a_{12}/(2\,a_{11}) & -a_{13}/(2\,a_{11}) \\ -a_{21}/(2\,a_{22}) & 1/2 & -a_{23}/(2\,a_{22}) \\ -a_{31}/(2\,a_{33}) & -a_{32}/(2\,a_{33}) & 1/2 \end{bmatrix} .$$

Logo,

$$||C||_\infty = max\left(1/2 + \frac{|a_{12}| + |a_{13}|}{2\,|a_{11}|}, 1/2 + \frac{|a_{21}| + |a_{23}|}{2\,|a_{22}|}, 1/2 + \frac{|a_{31}| + |a_{32}|}{2\,|a_{33}|}\right) .$$

Visto que A possui diagonal estritamente dominante por linhas o valor máximo em causa é inferior à unidade, ou seja, $||C||_\infty < 1$, pelo que o método é convergente, independentemente do vector inicial que se considere.

Sendo D a matriz diagonal contendo as entradas da diagonal principal de A, e u limite da sucessão definida pelo processo iterativo, resulta das respectivas fórmulas computacionais que

$$u = u + 1/2\,D^{-1}\,b - 1/2\,D^{-1}\,(A\,u) \Longleftrightarrow 0 = 1/2\,D^{-1}(b - A\,u) .$$

Por conseguinte, u é necessariamente solução do sistema dado.

2 a) A tabela em causa é

x_i	$-\pi/2$	0	$\pi/2$	π
$f(x_i)$	-1	0	1	0

Como $g(0) = 0 \Longleftrightarrow c = 0$, tem-se

$$g(x) = b\,x, \quad b \in \mathbb{R} .$$

Pretende-se minimizar

$$Q(b) = \sum_{i=0}^{3}(f(x_i) - g(x_i))^2 = \sum_{i=0}^{3}(b\,x_i - f(x_i))^2 .$$

Ora,

$$Q'(b) = 0 \Longleftrightarrow \sum_{i=0}^{3}(b\,x_i - f(x_i)) = 0 \Longleftrightarrow b = \frac{\sum_{i=0}^{3} f(x_i)}{\sum_{i=0}^{3} x_i} = 0 .$$

Por conseguinte, a melhor aproximação de mínimos quadrados de Ω é a função nula $g(x) = 0$.

2 b) Sendo $L_i(x)$ os elementos da base de Lagrange associada aos nós $\{-\pi/2, 0, \pi/2, \pi\}$, sabemos que

$$q(x) = (-1)\,L_0(x) + (1)\,L_2(x) .$$

Ora

$$L_0(x) = \frac{x\,(x - \pi/2)\,(x - \pi)}{(-\pi/2)\,(-\pi)\,(-3/2\,\pi)} = -4\frac{x\,(x - \pi/2)\,(x - \pi)}{3\,\pi^3},$$

$$L_2(x) = \frac{(x + \pi/2)\,x\,(x - \pi)}{(\pi)\,(\pi/2)\,(-\pi/2)} = -4\frac{x\,(x + \pi/2)\,(x - \pi)}{\pi^3} .$$

Assim,

$$q(x) = 8\frac{(\pi - x)\, x\, (x + \pi)}{3\,\pi^3} \ .$$

Como $f \in C^4([\pi/2, \pi])$ e $f^{(4)}(x) = \sin(x)$, para $x = 1/2$ sabe-se que existe pelo menos um valor $\theta \in (-\pi/2, \pi)$ tal que o erro de interpolação nesse ponto satisfaz a igualdade

$$f(1/2) - q(1/2) = f^{(4)}(\theta)/4!\,(1/2 + \pi/2)\,(1/2 - 0)\,(1/2 - \pi/2)\,(1/2 - \pi) \ .$$

Por conseguinte, a constante K tem por valor o produto dos 4 últimos factores na expressão anterior.

3 (a) Se a regra R for exacta para os monómios $1, x, x^2$ e x^3, sabe-se que será exacta para qualquer polinómio $p \in \mathcal{P}_3$. Ora,

$$R(1) = h \quad \text{e} \quad I(1) = \int_{-h/2}^{h/2} = h \ .$$

Como $y = x$ e $y = x^3$ são funções ímpares, a regra é trivialmente exacta para estas funções. Resta verificar o caso de $y = x^2$:

$$R(x^2) = \frac{h}{2}\left(2 \times \frac{h^2}{4}\right) - \frac{h^2}{12}\left(2 \times \frac{h}{2} + 2 \times \frac{h}{2}\right) = h^3/12$$

e

$I(x^2) = 2\int_0^{h/2} x^2\, dx = h^3/12$.

3 (b) Para $h = 1$ e $f(x) = e^x - 2$, obtém-se

$$R(f) = \frac{1}{2}(e^{-1/2} - 2 + e^{1/2} - 2) - \frac{1}{12}(e^{1/2} - e^{-1/2}) \simeq -0.95922 \ .$$

A regra de Newton-Cotes fechada é a de Simpson, com passo $1/2$. Logo,

$$S(f) = \frac{1}{6}(f(-1/2) + 4\, f(0) + f(1/2)) = 1/6\,(e^{-1/2} - 2 + 4\,(-1) + e^{1/2} - 2) \simeq -0.95746 \ .$$

Como $I(f) = \displaystyle\int_{-1/2}^{1/2} (e^x - 2)\, dx = e^{1/2} - e^{-1/2} - 2 \simeq -0.95780939$, conclui-se que o erro absoluto de $S(f)$ é inferior ao da regra $R(f)$.

4 (a) Para $f(x, y) = x\, \cos(x\, y)$ e $h = 1/2$, tem-se

$$f\,(x + h/2, y + h/2\, f(x, y)) = \left(x + \frac{1}{4}\right)\, \cos\left((x + \frac{1}{4})\,(y + 1/4(x\, \cos(x\, y))\right) \ .$$

Iniciando com $x_0 = 1$, bastam $N = 2$ passos para se aproximar $y(2)$:

$$\begin{cases} y_0 = y(1) = -1 \\ y_{i+1} = y_i + \dfrac{1}{2}\left\{(x_i + 1/4)\, \cos\left((x_i + 1/4)\,(y_i + \dfrac{x_i\, \cos(x_i\, y_i)}{4})\right)\right\}, \quad i = 0, 1 \ . \end{cases}$$

Assim,

$$y_1 = -1 + 1/2\ \{5/4\ \cos(5/4\,(-1 + \cos(-1)/4))\} \simeq -0.70606\ .$$

Para $x_1 = 3/2$, tem-se

$$y_2 = y_1 + 1/2\ \{7/4\ \cos(7/4\,(y_1 + 3/8\ \cos(3/2\,y_1)))\} \simeq -0.17197\ .$$

Donde, $y(2) \simeq -0.172$.

4 (b) Dado tratar-se de um método de segunda ordem, ele é exacto para funções y que sejam polinomiais de grau não superior a 2 . Por conseguinte o erro em causa deveria ser nulo.

A.2.24

Exame 11/01/2016 (Parte 1)

1. No sistema de vírgula flutuante $FP(10, 6, -20, 20)$ pretende-se calcular $g(10^{10})$, onde

$$g(x) = (\sqrt{x+1} - \sqrt{x-1})\,\sqrt{x}\ .$$

(a) Descreva, passo a passo, o algoritmo utilizado e diga que valor se obtém. Sabendo que o resultado exacto é um número positivo, como explica o resultado obtido?

(b) Proponha um algoritmo diferente que permita calcular $g(10^{10})$, no mesmo sistema VF, com um resultado mais preciso. Que valor se obtém nesse caso?
Sugestão: use a fórmula $\sqrt{a} - \sqrt{b} = \dfrac{a-b}{\sqrt{a}+\sqrt{b}}$.

2 (a) Considere a função $f(x) = x\,\sin x - 1$ no intervalo $I = [\pi/4, \pi/2]$. Mostre que f tem um único zero em I. Quantas iterações do método da bissecção são necessárias para determinar este zero com um erro absoluto inferior a 10^{-2} ?

2 (b) Determine β de modo a que o método iterativo

$$x_{n+1} = 1/\sin x_n\,, \qquad n = 0, 1, \ldots\,, \tag{A.5}$$

convirja para $z \in I$ qualquer que seja $x_0 \in [1, 1+\beta]$.

(c) Partindo de $x_0 = 1.1$ e aplicando o método (A.5), calcule uma aproximação do zero z de f com um erro absoluto inferior a 10^{-1}. Justifique a precisão obtida.

3) Considere o o sistema não linear

$$\begin{cases} x_1^2 + \alpha x_2 = 0 \\ \alpha x_1 + x_2^2 + \alpha x_3 = 0 \\ \alpha x_2 + x_3^2 = 1, \end{cases}$$

onde $\alpha \in \mathbb{R}$.

(a) Ao resolver este sistema pelo método de Newton, partindo da aproximação inicial $x^{(0)} = (1, 1, 1)$, somos conduzidos um sistema linear $Bx = d$, onde a matriz B é da forma

$$B = \begin{bmatrix} 2 & \alpha & 0 \\ \alpha & 2 & \alpha \\ 0 & \alpha & 2 \end{bmatrix}.$$

Explique como se obteve a matriz B e calcule o vector d.

(b) Supondo que se pretende resolver o sistema $Bx = d$, referido na alínea anterior, pelo método de Jacobi, mostre que este método converge se e só se $|\alpha| < \sqrt{2}$.

(c) Sabendo que

$$B^{-1} = \frac{1}{8 - 4\alpha^2} \begin{bmatrix} 4 - \alpha^2 & -2\alpha & \alpha^2 \\ -2\alpha & 4 & -2\alpha \\ \alpha^2 & -2\alpha & 4 - \alpha^2 \end{bmatrix},$$

diga para que valores de α a matriz B é mal condicionada.

Resolução

1 a) Algoritmo para calcular $g(x)$:
Passo 1. $z_1 = 1 + x$;
Passo 2. $z_2 = \sqrt{z_1}$;
Passo 3. $z_3 = x - 1$;
Passo 4. $z_4 = \sqrt{z_3}$;
Passo 5. $z_5 = z_2 - z_4$;
Passo 6. $z_6 = \sqrt{x}$;
Passo 7. $z = z_5 * z_6$;

Para $x = 10^{10}$, e atendendo a que $fl(x + 1) = fl(x - 1) = fl(x) = 0.1 \times 10^{11}$, temos:

$$z_1 = 0.1 \times 10^{11};\ z_2 = 0.1 \times 10^{6};\ z_3 = 0.1 \times 10^{11};\ z_4 = 0.1 \times 10^{6}$$
$$z_5 = 0;\ z_6 = 0.1 \times 10^{6};\ z = 0\ .$$

Ou seja, o resultado final é 0. Isto acontece porque no passo 5, ao subtrair dois valores muito próximos entre si, ocorre *cancelamento subtractivo*, introduzindo um grande erro relativo nos cálculos. Por outras palavras, o algoritmo apresentado é instável para valores elevados de x.

1 b) Utilizando a sugestão, temos

$$g(x) = (\sqrt{x+1} - \sqrt{x-1})\sqrt{x} = \frac{2}{\sqrt{x+1} + \sqrt{x-1}}\sqrt{x} \qquad (*)$$

Com base na fórmula $(*)$, considere-se o algoritmo:

Passo 1. $w_1 = 1 + x$;
Passo 2. $w_2 = \sqrt{w_1}$;
Passo 3. $w_3 = x - 1$;
Passo 4. $w_4 = \sqrt{w_3}$;
Passo 5. $w_5 = w_2 + w_4$;
Passo 6. $w_6 = 2/w5$;
Passo 7. $w_7 = \sqrt{x}$;
Passo 8. $w = w_6 * w_7$.

Este algoritmo é preferível, no caso de valores elevados de x, porque evita o cancelamento subtractivo e, deste modo, diminui muito significativamente o efeito dos erros de arredondamento. Usando o novo algoritmo no caso de $x = 10^{10}$, obtém-se

$$\begin{aligned} &w_1 = 0.1 \times 10^{11};\ w_2 = 0.1 \times 10^6 \\ &w_3 = 0.1 \times 10^{11};\ w_4 = 0.1 \times 10^6 \\ &w_5 = 0.2 \times 10^6;\ w_6 = 0.1 \times 10^{-4} \\ &w_7 = 0.1 \times 10^6;\ w = 0.1 \times 10 = 1\ . \end{aligned}$$

Ou seja, o resultado é 1, muito próximo do valor exacto de $g(10^{10})$.

2 a) A função f é infinitamente diferenciável em $\mathbb{R}$. Por outro lado, temos $f(\pi/4) = -1 + \dfrac{\pi}{4\sqrt{2}} \simeq -0.446$, $f(\pi/2) = -1 + \dfrac{\pi}{2} \simeq 0.571$, pelo que existe, pelo menos, um zero de f em I . Além disso, $f'(x) = x\cos x + \sin x > 0$ em I, logo f tem um único zero no intervalo. Quanto ao número de iterações do método da bissecção, necessárias para determinar este zero com um erro absoluto inferior a 10^{-2}, basta calcular $\log_2(\dfrac{\pi/2 - \pi/4}{0.01}) \simeq 6.3$. Logo, o número de iterações é 7 .

2 b) Em primeiro lugar, sendo a função iteradora $g(x) = 1/\sin(x)$, tem-se que

$$f(x) = 0 \quad \Longleftrightarrow \quad x\sin(x) = 1 \quad \Longleftrightarrow \quad x = g(x)\ .$$

Logo, as raizes de f são pontos fixos de g . Consideremos, por exemplo, $\beta = 0.2$. Verifiquemos se as condições do teorema do ponto fixo estão satisfeitas em $I = [1, 1.2]$.

(i) Temos

$$g(1) = 1/\sin(1) = 1.184 \in I, \qquad g(1.2) = 1/\sin(1.2) = 1.073 \in I\ .$$

Além disso, $g'(x) = -\cos(x)/\sin(x)^2 < 0$, em I, logo g é monótona no intervalo . Por conseguinte, $g(I) \subset I$.

(ii) Já vimos que $g'(x) < 0$ em I. Verifiquemos se $g'(x)$ é monótona. Para isso, determinemos $g''(x) = \dfrac{\sin^3 x + 2\sin x \cos^2 x}{\sin^4 x} > 0$, pelo que g' é monótona crescente em I. Assim,

$$\max_{x\in I} |g'(x)| = |g'(1)| = 0.76306\ .$$

Ou seja, $L = 0.76306 < 1$.

Pelo teorema do ponto fixo, o método iterativo converge no intervalo $I = [1, 1.2]$, isto é, com $\beta = 0.2$. (Do mesmo modo se poderia demonstrar para outros valores de β, tais que $0.19 \leq \beta \leq 1.49$.)

2 c) Sendo $x_0 = 1.1$, temos $x_1 = x_0/\sin(x_0) \simeq 1.122$.
$|x_1 - z| \leq \dfrac{L}{1-L}|x_1 - x_0| \simeq 0.071$. A aproximação x_1 tem erro inferior a 10^{-1}, isto é, satisfaz a condição imposta.

3 a). Este sistema pode ser escrito na forma $(F_1, F_2, F_3) = (0, 0, 0)$, onde

$$\begin{array}{ll} F_1(x_1, x_2, x_3) & = x_1^2 + \alpha\, x_2 \\ F_2(x_1, x_2, x_3) & = x_2^2 + \alpha\, x_1 + \alpha\, x_3 \\ F_3(x_1, x_2, x_3) & = \alpha x_2 + x_3^2 - 1\ . \end{array}$$

Sendo assim, a matriz jacobiana de F tem a forma

$$J_F(x_1, x_2, x_3) = \left[\begin{array}{ccc} 2\,x_1 & \alpha & 0 \\ \alpha & 2\,x_2 & \alpha \\ 0 & \alpha & 2\,x_3 \end{array}\right]\ .$$

Uma vez que a aproximação inicial é $x^{(0)} = (1, 1, 1)$, a matriz do sistema linear para o método de Newton obtém-se calculando $J_F(1, 1, 1)$, o que nos dá a matriz B .

Para obter o vector d, basta calcular

$$d = -(F_1(1, 1, 1)), F_2(1, 1, 1)), F_3(1, 1, 1)) = -(1 + \alpha, 1 + 2\,\alpha, \alpha)\ .$$

3 b) Em primeiro lugar, determinemos a matriz de iteração do método de Jacobi para este sistema linear:

$$C_J = -D^{-1}(L + U) = \left[\begin{array}{ccc} 0 & -\alpha/2 & 0 \\ -\alpha/2 & 0 & -\alpha/2 \\ 0 & -\alpha/2 & 0 \end{array}\right].$$

O polinómio característico desta matriz é

$$\det(C_J - \lambda I) = -\lambda^3 + \frac{\alpha^2}{2}\lambda,$$

e os seus valores próprios são $\lambda_1 = 0$, $\lambda_{2,3} = \pm\dfrac{\alpha}{\sqrt{2}}$. Logo, o seu raio espectral é $\rho(C_J) = |\dfrac{\alpha}{\sqrt{2}}|$. A condição necessária e suficiente de convergência do método de Jacobi é $\rho(C_J) < 1$, o que equivale a $|\alpha| < \sqrt{2}$.

3 c) Escolhendo, por exemplo, a norma do máximo, O número de condição de B é dado por $\mathrm{cond}(B) = \|B\|_\infty \|B^{-1}\|_\infty$. Neste caso, temos

$$\|B\|_\infty = \max\left(2 + |\alpha|, 2 + 2|\alpha|\right) = 2 + 2|\alpha|,$$

$$\|B^{-1}\|_\infty = \frac{1}{|8 - 4\alpha^2|}\max\left(|4 - \alpha^2| + 2|\alpha| + \alpha^2, 4 + 4|\alpha|\right)\ .$$

O número de condição pode tomar valores altos quando o denominador de $\|B^{-1}\|_\infty$ for próximo de 0 . Isto acontece se $8-4\alpha^2 \approx 0$, isto é , $\alpha \approx \pm\sqrt{2}$. Neste caso, temos

$$\lim_{\alpha\to\pm\sqrt{2}} \|B\|_\infty = \lim_{\alpha\to\pm\sqrt{2}} 2+2|\alpha| = 2+2\sqrt{2}\ .$$

$$\lim_{\alpha\to\pm\sqrt{2}} \|B^{-1}\|_\infty = \lim_{\alpha\to\pm\sqrt{2}} \frac{4+|4\alpha|}{8-4\alpha^2} = \infty\ .$$

Por conseguinte, temos que

$$\lim_{\alpha\to\pm\sqrt{2}} cond(B) = \infty,$$

o que significa que *B é mal condicionada para valores de α próximos de $\pm\sqrt{2}$* . Vejamos o que acontece quando $\alpha \to \pm\infty$. Como

$$\lim_{\alpha\to\pm\infty} \|B\|_\infty = \lim_{\alpha\to\pm\infty} 2+2|\alpha| = \infty,$$

$$\lim_{\alpha\to\pm\infty} \|B^{-1}\|_\infty = \lim_{\alpha\to\pm\sqrt{2}} \frac{2|\alpha|+2\alpha^2-4}{4\alpha^2-8} = \frac{1}{2}\ .$$

Tem-se

$$\lim_{\alpha\to\pm\infty} cond(B) = \infty,$$

ou seja, *B é mal condicionada para valores elevados de $|\alpha|$.*

Para outros valores de α nem $\|B\|_\infty$, nem $\|B^{-1}\|_\infty$ podem tomar valores muito elevados, pelo que a matriz é bem condicionada.

A.2.25

Exame 11/01/2016 (Parte 2)

1) Considere a seguinte tabela de diferenças de uma função f:

x_i	f_i			
2	0.7			
		3.9		
5	y		z	
		9.1		1/6
6	21.5		22/15	
		37/6		
v	3			

(a) Calcule os valores de v, y e z nesta tabela e determine o polinómio que interpola f nos 4 pontos considerados. (Caso não a resolva esta alínea, nas alíneas seguintes considere $v=3$, $y=12.4$ e $z=1.3$.)

(b) Sabendo que o polinómio que interpola f nos pontos $1,2,5,6$ é $P_3(x)=(x^3-1)/10$, e com base na alínea anterior, determine o polinómio P_4 que interpola f em $1,2,5,6,v$.

(c) Obtenha um valor aproximado de $\int_2^6 f(x)dx$, usando a regra dos trapézios e os 4 valores tabelados de f .
Sugestão: tenha em conta que

$$\int_2^6 f(x)dx = \int_2^v f(x)dx + \int_v^5 f(x)dx + \int_5^6 f(x)dx \ .$$

(d) Obtenha uma nova aproximação de $\int_2^6 f(x)dx$, usando uma regra da forma

$$Q(f) = Af(2) + Bf(5) + Cf(6)$$

e determinando os pesos A, B, C pelo método dos coeficientes indeterminados.

(e) Diga, justificando, qual das regras consideradas nas duas alíneas anteriores tem o grau de precisão mais elevado.

2) Considere uma curva no plano cartesiano que passa pelos pontos: $(-\pi, 0)$, $(-\pi/2, 1)$, $(0, -1)$, $(\pi/2, 0)$ e $(\pi, 1)$. Determine a função da forma

$$g(x) = a + b\sin(x) + c\cos(x)$$

que melhor aproxima esta curva, no sentido dos mínimos quadrados.

3) Considere a equação diferencial

$$y'(x) = -2\,y(x) + \frac{a}{1 + y(x)^2},$$

com a condição inicial $y(0) = y_0$.

(a) No caso de $a = 1$ e $y_0 = -1$ obtenha um valor aproximado de $y(0.2)$, efectuando dois passos do método de Euler.

(b) Seja $a = 0$. Representando por y_i o valor aproximado de $y(x_i)$, obtido pelo método de Euler, onde $x_i = i\,h$, e sabendo que a solução exacta deste problema é $y(x) = y_0\,e^{-2x}$, diga que condição deve satisfazer o valor de h para que o erro da aproximação y_i satisfaça

$$\lim_{i\to\infty} |y(x_i) - y_i| = 0 \ . \tag{A.6}$$

Sugestão: mostre que é válida a fórmula

$$y_i = y_0(1 - 2h)^i, \quad i = 1, 2, \ldots \ . \tag{A.7}$$

Resolução

1 a) Para calcular v:

$$\frac{3 - 21.5}{v - 6} = \frac{37}{6},$$

donde $v = 3$. Para calcular y :

$$\frac{21.5 - y}{6 - 5} = 9.1,$$

pelo que $y = 12.4$. Finalmente, para calcular z,

$$z = \frac{9.1 - 3.9}{6 - 2} = 1.3 \ .$$

O polinómio que interpola f nos pontos 2,3, 5 e 6, de acordo com a fórmula interpoladora de Newton, tem a forma

$$\begin{aligned} Q_3(x) &= 0.7 + 3.9(x-2) + 1.3(x-2)(x-5) + \tfrac{1}{6}(x-2)(x-5)(x-6) \\ &= -4.1 + 3.4667x - 0.8667x^2 + \frac{1}{6}x^3 \ . \end{aligned}$$

1 b) Da forma de $P_3(x)$ resulta que $f[1,2,5,6] = 1/10$ (coeficiente de x^3) . Por outro lado, da tabela resulta que $f[2,5,6,3] = 1/6$. Destas duas igualdades conclui-se que

$$f[1,2,5,6,3] = (f[1,2,5,6] - f[2,5,6,3])/(1-3) = 1/30 \ .$$

Usando mais uma vez a fórmula interpoladora de Newton, obtém-se

$$\begin{aligned} P_4(x) &= Q_3(x) + \tfrac{1}{30}(x-2)(x-5)(x-6)(x-3) \\ &= 1.9 - 3.733x + 2.1667x^2 - 11/30x^3 + x^4/30 \ . \end{aligned}$$

1 c) Não se pode utilizar a regra dos trapézios composta, porque o espaçamento entre os pontos da tabela não é constante. Temos $x_0 = 2$, $x_1 = v = 3$, $x_2 = 5$, $x_3 = 6$. Assim, usando a sugestão, aplica-se a regra dos trapézios simples a cada subintervalo.

$$\int_2^3 f(x)dx \approx (f(2) + f(3))/2 * (3-2) = (0.7+3)/2 = 1.85 \ .$$

$$\int_3^5 f(x)dx \approx (f(3) + f(5))/2 * (5-3) = 3 + 12.4 = 15.4 \ .$$

$$\int_5^6 f(x)dx \approx (f(5) + f(6))/2 * (6-5) = (12.4 + 21.5)/2 = 16.95 \ .$$

Reunindo os 3 resultados, obtém-se

$$\int_2^6 f(x)dx \approx 1.85 + 15.4 + 16.95 = 34.2 \ .$$

1 d) Para aplicar o método dos coeficientes indeterminados, com os pontos $x_0 = 2$, $x_1 = 5$, $x_2 = 6$, deverá resolver-se o seguinte sistema de equações lineares:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 5 & 6 \\ 2^2 & 5^2 & 6^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 6-2 \\ (6^2-2^2)/2 \\ (6^3-2^3)/3 \end{bmatrix} .$$

A solução deste sistema é $A = 10/9$,$B = 32/9$, $C = -2/3$. Substituindo na regra de quadratura, obtém-se

$$\begin{aligned} Q(f) &= Af(2) + Bf(5) + Cf(6) \\ &= 10/9 * 0.7 + 32/9 * 12.4 - 2/3 * 21.5 = 30.533 \ . \end{aligned}$$

1 e) A regra dos trapézios tem grau 1. Quanto à regra $Q(f)$, obtida pelo método dos coeficientes indeterminados, é exacta para qualquer polinómio de grau menor ou igual a 2, pelo que o seu grau é pelo menos 2. Por conseguinte, esta última fórmula tem um grau de precisão mais elevado.

2) De acordo com o enunciado temos um ajustamento linear com as seguintes funções de base:

$$\phi_0(x) = 1, \qquad \phi_1(x) = \cos(x), \qquad \phi_2(x) = \sin(x) \ .$$

As abcissas dos pontos onde a curva é aproximada são: $x_0 = -\pi$, $x_1 = -\pi/2$, $x_2 = 0$, $x_3 = \pi/2$ e $x_4 = \pi$.
Usando notação vectorial, temos assim

$$\begin{aligned} \phi_0 &= (1,1,1,1,1)^T \\ \phi_1 &= (-1,0,1,0,-1)^T \\ \phi_2 &= (0,-1,0,1,0)^T \\ f &= (0,1,-1,0,1)^T \ . \end{aligned}$$

Logo, a matriz do sistema normal tem a forma

$$\begin{bmatrix} (\phi_0,\phi_0) & \phi_0,\phi_1) & (\phi_0,\phi_2) \\ (\phi_1,\phi_0) & \phi_1,\phi_1) & (\phi_1,\phi_2) \\ (\phi_2,\phi_0) & (\phi_2,\phi_1) & (\phi_2,\phi_2) \end{bmatrix} = \begin{bmatrix} 5 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

enquanto o segundo membro do sistema é

$$((\phi_0,f),(\phi_1,f),(\phi_2,f))^T = (1,-2,-1)^T \ .$$

Finalmente, resolvendo o sistema, obtém-se $a = 1/14, b = -9/14, c = -1/2$. Logo, a solução é $g(x) = 1/14 - 9/14\ \cos(x) - 1/2\ \sin(x)$.

3 a) Visto que $x_0 = 0$ e se pretende efectuar dois passos, então o comprimento do passo deve ser $h = 0.1$ e os pontos onde se aproxima a solução são $x_1 = 0.1$ e $x_2 = 0.2$.
Primeiro passo:

$$y_1 = y_0 + h\left(-2y_0 + \frac{1}{1+y_0^2}\right) = -1 + h\left(2 + \frac{1}{2}\right) = -0.75 \ .$$

Segundo passo:

$$y_2 = y_1 + h\left(-2y_1 + \frac{1}{1+y_1^2}\right) = -0.75 + h\left(1.5 + \frac{1}{1+0.75^2}\right) = -0.536 \ .$$

3 b) Comecemos por provar que $y_i = y_0(1-2h)^i$. Para isso, verifiquemos primeiro que $y_1 = y_0 + h\,(-2y_0)$, ou seja, $y_1 = y_0(1-2h)$. Logo, a fórmula (A.7) é válida para $i = 1$.

Suponhamos agora que a fórmula é válida para $i = n$, isto é, $y_n = y_0\,(1-2h)^n$. Nesse caso, temos

$$y_{n+1} = y_n + h(-2y_n) = y_0(1-2h)^n - 2hy_0(1-2h)^n = y_0(1-2h)^{n+1}.$$

A fórmula (A.7) fica assim provada por indução.

Note-se que a solução exacta satisfaz $\lim_{i\to\infty} y(x_i) = 0$. Por isso, para que se verifique (A.6) basta que $\lim_{i\to\infty} y_i = 0$.

Uma vez que a fórmula (A.6) é válida, a condição necessária e suficiente para que $\lim_{i\to\infty} y_i = 0$ é que $|1-2h| < 1$. Isto verifica-se se e só se $h < 1$.

A.2.26

Teste de 25 de Maio de 2016

1) Seja h um número real positivo. Considere o sistema linear, com solução $x = (1,1,1)$,

$$\begin{cases} -2\,h\,x_1 + x_3 & = 1 - 2\,h \\ h\,x_1 + x_2 & = 1 + h \\ -h\,x_1 + 2\,h\,x_3 & = h, \end{cases}$$

(a) Calcule a matriz de iteração do método de Gauss-Seidel aplicado ao sistema. Mostre que este método é convergente para a solução x, qualquer que seja o valor $h > 1/4$ que
[1.0] considere. Justifique.

[1.0] **(b)** Fazendo $h = 1$ e $x^{(0)} = (-1,0,1)$, obtenha a primeira iterada $x^{(1)}$ do método referido na alínea anterior e calcule $||x^{(1)} - x||_1$. Apresente todos os cálculos que efectuar.

2) Considere os valores

x_i	0	$-1/2$	1
$f(x_i)$	0	-0.4431	1.693

[1.0] **(a)** Calcule o polinómio interpolador de Newton para a tabela dada. Apresente todos os cálculos.

[1.0] **(b)** Admita que os valores tabelados se referem a uma função $f(x) = \ln(x+1) + q(x)$, onde $q(x)$ representa um certo polinómio de grau 1. Se no intervalo $A = [-1/2, 0]$ aproximar a função f por um polinómio interpolador nos nós $-1/2$ e 0, sem calcular $q(x)$ obtenha um majorante do erro absoluto de interpolação nesse intervalo. Justifique.

[1.5] **(c)** Pretende-se calcular a melhor aproximação de mínimos quadrados dos valores tabelados por funções aproximantes do tipo $h(x) = a + b\,\sin(x)$, onde $a, b \in \mathbb{R}$. Depois de escrever a função a minimizar, deverá obter todas as entradas do sistema de equações normais usando arredondamento simétrico para duas casas decimais, e escrever tal sistema. Não é necessário calcular a solução.

[1.0] **3 a)** Aplique convenientemente a regra dos trapézios para aproximar $\int_{-1/2}^{1} f(x)dx$,

conhecidos os valores da tabela dada em **2)**.

[1.0] **3 b)** Seja $f(x) = \ln(x)$, $x \in [1,2]$. Obtenha uma regra de quadratura da forma $Q(f) = \alpha\, f(1) + \beta\, f(2)$, mediante aplicação do método dos coeficientes indeterminados. Qual é a designação habitual da regra que determinou? Justifique.

4) No intervalo $I = [0, \pi]$, considere o problema

$$y'(t) = -2\,\sin(t)\, y^2(t), \quad y(0) = 1\ .$$

(a) Mostre que o problema tem solução, qualquer que seja o valor inicial $y(0)$ que considere. Efectuando cálculos exactos, obtenha uma aproximação de $y(\pi/2)$ mediante dois passos do método de Euler explícito. Designe o valor que calculou por $\tilde{y}$. **[1.5]**

(b) Admita que $\lim_{t\to\pi/2}[y(t)] = 1/3$. Mostre que o erro absoluto de $\tilde{y}$ é menor do que $1/2$. Justifique. **[1.0]**

Resolução

1 (a) O sistema dado é equivalente a

$$\left\{\begin{array}{l} x_1 = \dfrac{-1+2\,h+x_3}{2\,h} \\ x_2 = 1+h-h\,x_1 = 1+h-\dfrac{-1+2\,h+x_3}{2} = \dfrac{3-x_3}{2} \\ x_3 = \dfrac{h+h\,x_1}{2\,h} = \dfrac{h+\frac{-1+2\,h+x_3}{2}}{2\,h} = \dfrac{4\,h-1+x_3}{4\,h} \end{array}\right. \qquad (**)$$

Levando em conta as igualdades anteriores, a matriz de iteração C_{GS} do método de Gauss-Seidel aplicado ao sistema, escreve-se:

$$C_{GS} = \left[\begin{array}{ccc} 0 & 0 & 1/(2\,h) \\ 0 & 0 & -1/2 \\ 0 & 0 & 1/(4\,h) \end{array}\right].$$

Logo, $\rho(C_{GS}) = 1/(4\,h) < 1$, se e só se $h > 1/4$.
Como verificação, note-se que

$$(D+L)^{-1} = \left[\begin{array}{ccc} -2\,h & 0 & 0 \\ h & 1 & 0 \\ -h & 0 & 2\,h \end{array}\right]^{-1} = \left[\begin{array}{ccc} -1/(2\,h) & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/(4\,h) & 0 & 1/(2\,h) \end{array}\right],$$

donde

$$C_{GS} = -(D+L)^{-1}\,U = -\left[\begin{array}{ccc} -1/(2\,h) & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/(4\,h) & 0 & 1/(2\,h) \end{array}\right]\left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{ccc} 0 & 0 & 1/(2\,h) \\ 0 & 0 & -1/2 \\ 0 & 0 & 1/(4\,h) \end{array}\right].$$

1 (b) Para $h = 1$, atendendo às expressões $(**)$, as fórmulas de iteração são:

$$\left\{\begin{array}{l} x_1^{(k+1)} = \dfrac{1+x_3^{(k)}}{2} \\ x_2^{(k+1)} = \dfrac{3-x_3^{(k)}}{2}, \qquad k = 0, 1, \ldots \\ x_3^{(k+1)} = \dfrac{3+x_3^{(k)}}{4} \end{array}\right.$$

Fazendo $x^{(0)} = (-1, 0, 1)$, obtém-se

$$\begin{aligned} x^{(1)} &= (1, 1, 1) \\ x - x^{(1)} &= (0, 0, 0) \ . \end{aligned}$$

Assim, $||x^{(1)} - x||_1 = 0$.

2 (a) Dado que

$$\begin{aligned} f[0, -1/2] &= \frac{-0.4431}{-0.5} = 0.8862 \\ f[-1/2, 1] &= \frac{1.693 + 0.4431}{1.5} \simeq 1.42407 \\ f[0, -1/2, 1] &= \frac{1.42407 - 0.8862}{1} = 0.53787 \ . \end{aligned}$$

o polinómio interpolador para a tabela dada é

$$p_2(x) \simeq 0.8862\, x + 0.53787\, x\, (x + 1/2) \ .$$

2 (b) Seja p_1 o polinómio interpolador nos nós $\{-1/2, 0\}$ e $E = max_{x\in A}|f(x) - p_1(x)|$. Tem-se que

$$E \leq \frac{max_{x\in A}|f^{(2)}(x)|}{2} \times max_{x\in A}\, |(x + 1/2)\, x| \ .$$

Ora,

$$\begin{aligned} f(x) &= \ln(x + 1) + q(x), \quad \text{onde} \quad q \in \mathcal{P}_1 \\ f'(x) &= \frac{1}{x + 1} + q'(x) \\ f^{(2)}(x) &= -\frac{1}{(x + 1)^2} + 0 \ . \end{aligned}$$

Assim,

$$max_{x\in[-1/2,0]}|f^{(2)}(x)| = max_{x\in[-1/2,0]}\frac{1}{(x + 1)^2} = 4 \ .$$

Sendo $w_2(x) = (x + 1/2)\, x$, resulta $w_2'(x) = 2\, x + 1/2 = 0$ se e só se $x = -1/4$. Logo, $w_2(-1/4) = 1/16 - 1/8 = -1/16$. Por conseguinte,

$$E \leq 4/2 \times 1/16 = 1/8 \ .$$

2 (c) Seja $h(x) = a\, \phi_0(x) + b\, \phi_1(x)$, $f = (0, -0.44, 1.7)$. A função a minimizar é $F(a, b) = \sum_{i=0}^{i=2}(h(x_i) - f(x_i))^2, \quad a, b \in \mathbb{R}$.
Tem-se,

$$\begin{aligned} &\phi_0(x) = 1 \Longrightarrow \phi_0 = (1, 1, 1) \\ &\phi_1(x) = \sin(x) \Longrightarrow \phi_1 = (0, \sin(-1/2), sin(1)) \ . \end{aligned}$$

Por conseguinte,

$$\begin{aligned}
&< \phi_0, \phi_0 >= 3\\
&< \phi_0, \phi_1 >= \sin(-1/2) + \sin(1) = 0.3620\cdots\\
&< \phi_1, \phi_1 >= \sin^2(-1/2) + \sin^2(1) = 0.9379\cdots\\
&< \phi_0, f >= 0 - 0.44 + 1.7 = 1.2499\\
&< \phi_1, f >= -0.44\,\sin(-1/2) + 1.7\,\sin(1) = 1.6370\cdots .
\end{aligned}$$

O sistema de equações normais a resolver é

$$\left[\begin{array}{cc} 3 & 0.36 \\ 0.36 & 0.94 \end{array}\right] \left[\begin{array}{c} a \\ b \end{array}\right] = \left[\begin{array}{c} 1.25 \\ 1.64 \end{array}\right] .$$

3 (a) Designando por $T[\alpha_0, \alpha_1]$ a regra dos trapézios (simples) aplicada ao intervalo $[\alpha_0, \alpha_1]$, tem-se

$$I(f) = \int_{-1/2}^{1} f(x)\,dx \simeq T[-1/2, 0] + T[0, 1],$$

isto é,

$$\begin{aligned}
I(f) &= \frac{1}{4}\,(f(-1/2) + f(0)) + \frac{1}{2}\,(f(0) + f(1))\\
&= 0.25\,(-0.4431 + 0) + 0.5\,(0 + 1.693) = 0.735725\,.
\end{aligned}$$

3 (b) No intervalo em causa, a regra é exacta para polinómios de grau ≤ 1 se e só se é exacta para os polinómios 1 e x, isto é, $Q(1) = \int_1^2 dx = 1$ e $Q(x) = \int_1^2 x\,dx = 3/2$. Donde,

$$\left\{\begin{array}{l} \alpha + \beta = 1 \\ \alpha + 2\,\beta = 3/2\ . \end{array}\right.$$

O sistema anterior tem solução (única) $\alpha = \beta = 1/2$. Assim, a regra $Q(f) = 1/2\,(\ln(1) + \ln(2)) = \ln(2)/2$ é a regra dos trapézios simples aplicada à função f, no intervalo $[1, 2]$.

4 (a) Em $D = [0, \pi] \times \mathbb{R}$ as funções $f(t, y) = -2\sin(t)\,y^2$, $\partial f(t, y)/\partial t = -2\cos(t)\,y$ e $\partial f(t, y)/\partial y = -4\sin(t)\,y$ são contínuas. Sabe-se que tais condições são suficientes para garantir existência e unicidade de solução para $y'(t) = f(t, y(t))$, com $y(t_0) = y_0$, $t_0 \in I = [0, \pi]$ e $y_0 \in \mathbb{R}$. Por conseguinte, o problema de valor inicial considerado tem solução.

Sendo $t_0 = 0$ e $N = 2$ passos, deverá fazer-se $h = \dfrac{\pi/2 - 0}{N} = \pi/4$. Assim, para $y_0 = 1$, os dois passos do método de Euler explícito calculam-se mediante as fórmulas

$$y_{i+1} = y_i - 2\,h\sin(t_i)\,y_i^2 = y_i - (\pi/2)\,\sin(t_i)\,y_i^2, \qquad i = 0, 1\ .$$

Ou seja,

$$\begin{aligned}
y_1 &= 1 - (\pi/2)\sin(0) \times 1 = 1\\
y_2 &= 1 - (\pi/2)\sin(\pi/4) \times 1 = 1 - \frac{\sqrt{2}\,\pi}{4} \simeq -0.110721\ .
\end{aligned}$$

4 (b) Uma vez que a solução y é função contínua no intervalo I tem-se $y(\pi/2) = \lim_{t \to \pi/2} y(t) = 1/3$. Por conseguinte o erro do valor $\tilde{y}$ é

$$y(\pi/2) - \tilde{y} = 1/3 - \left(1 - \frac{\sqrt{2}\,\pi}{4}\right) \simeq 0.44 < 1/2\ .$$

A.2.27

Exame de 30 de Junho 2016 (Parte 1)

1) Considere os números reais $A = \sqrt{\alpha + 10.1}$, $B = \sqrt{\alpha + 10.2}$ e o sistema de ponto flutuante $FP(10, 3, -30, 30)$ com arredondamento simétrico.

(a) Tome para α o primeiro dígito decimal do seu número de aluno (por exemplo, $\alpha = 7$, número de aluno 78500). Calcule $W = A - B$ no referido sistema, bem como o erro relativo do valor que obteve. **[1.0]**

(b) Escreva uma expressão alternativa para W, cujo algoritmo seja numericamente estável. Justifique. (Não é necessário efectuar cálculos numéricos). **[1.0]**

2) A fim de obter o inverso aritmético do número $c > 0$, ou seja $1/c$, recorreu-se ao método iterativo

$$x_{k+1} = x_k\,(2 - c\,x_k), \quad k = 0, 1, \ldots$$

(a) Para $c = 5$ e $x_0 = 1/4$, calcule x_2 e o respectivo erro absoluto, efectuando cálculos exactos. **[1.0]**

(b) Verifique que em $\mathbb{R}$ existe apenas um ponto fixo (não nulo) do processo iterativo, o qual é solução do problema em causa. Como classifica um tal ponto fixo? Justifique. **[1.0]**

(c) Poderá afirmar que, uma vez iniciado o processo iterativo em $x_0 \in \mathbb{R}$, se houver convergência do processo a mesma é supralinear? Justifique. **[1.5]**

(d) Mostre que se iniciar o processo com $x_0 > 2/c$ o método não converge. Indique um valor inicial para o qual possa garantir convergência monótona para o valor $1/c$. Justifique. **[1.5]**

3) Para se aproximar a solução do sistema

$$\begin{cases} x_1^2 - x_2 + x_3^2 & = \alpha \\ -x_1 + x_2 - x_3 & = \beta \\ 2\,x_1 - x_2 + 3\,x_3 & = \gamma, \end{cases}$$

mediante aplicação do método de Newton, utilizou-se inicialmente $x^{(0)} = (1, 0, 1)^T$ a fim de se calcular a primeira iterada $x^{(1)}$, mediante resolução de um certo sistema linear $A\,u = b$.

(a) Efectuando cálculos exactos obtenha A^{-1} e mostre que $cond_1(A) < 25$. **[1.5]**

(b) Admitindo que $b = (1, 0, 1)^T$, calcule os valores de α, β e γ . **[0.5]**

(c) Sabe-se que o método de Gauss-Seidel converge para a solução de $A\,u = b$, se A é matriz simétrica definida positiva. Diga, justificando, se poderá assegurar a convergência deste método para $u = A^{-1}\,b$, caso inicie o processo com $x^{(0)} = b$. **[1.0]**

Resolução

1 (a) Por exemplo, para $\alpha = 7$,

$A = \sqrt{\alpha + 10.1} = \sqrt{17.1} = 4.13521\cdots \quad B = \sqrt{\alpha + 10.2} = \sqrt{17.2} = 4.14728\cdots$
$W = A - B = -0.01207\cdots$
$\bar{A} = fl(A) = 4.14 = 0.414 \times 10^1 \quad \bar{B} = fl(B) = 4.15 = 0.415 \times 10^1$
$\bar{W} = fl(\bar{A} - \bar{B}) = fl(-0.001 \times 10^1) = -0.01 = -0.100 \times 10^{-1}$.

Logo,

$$\delta_{\bar{W}} = \frac{W - \bar{W}}{W} = \frac{-0.00207\cdots}{-0.01207\cdots} \simeq 0.172 = 17.2\,\% .$$

1 (b) Dado que $A^2 - B^2 = (A - B)\,(A + B)$, resulta

$$W = A - B = \frac{A^2 - B^2}{A + B} = \frac{17.1 - 17.2}{\sqrt{17.1} + \sqrt{17.2}} .$$

O cálculo de $fl(W)$ minora o efeito de cancelamento subtractivo para os valores em questão.

2 (a)

$$\begin{array}{l} x_0 = 1/4; \quad x_{k+1} = x_k(2 - 5\,x_k), \quad k = 0, 1 \ . \\ x_1 = 1/4(2 - 5/4) = 3/16 \\ x_2 = 3(16\,(2 - 15/16) = 51/256 \Longrightarrow e_2 = 1/5 - x_2 = 1/1280 \ . \end{array}$$

2 (b) A solução do problema é o número inverso de c, $\tilde{x} = 1/c$. Dado que

$$x = x\,(2 - c\,x)\ \forall x \in \mathbb{R} \quad \text{se e só se} \quad x = 0 \text{ ou } 2 - c\,x = 1 \Longleftrightarrow x = 1/c,$$

conclui-se que a função iteradora g possui um só ponto fixo não nulo em $\mathbb{R}$. Uma vez que $g'(x) = 2 - 2\,c\,x = 2\,(1 - c\,x)$ e $g'(\tilde{x}) = 0$, tal ponto fixo é superatractor.

2 (c) Como para o ponto fixo $\tilde{x} = 0$ se tem $g'(0) = 2$, este ponto fixo é repulsor, logo o processo não pode convergir para este ponto. Quanto a $\tilde{x} = 1/c$, dado que $g'(\tilde{x}) = 0$, uma vez escolhido x_0 suficientemente próximo de $\tilde{x}$ o processo converge supralinearmente para $\tilde{x}$. Visto que a função g não tem outros pontos fixos além de 0 e $1/c$, se o processo não converge para um desses pontos necessariamente é divergente.

2 (d) Como sabemos, atendendo a que $g(x) = x\,(2 - c\,x)$, o único ponto fixo positivo de g é o ponto $\tilde{x} = 1/c$. Além disso, a função g é quadrática satisfazendo a desigualdade

$$g(x) < 0, \quad \forall x > 2/c \ .$$

Assim, se $x_0 > 2/c$, tem-se $x_1 < 0$. Por outro lado, $g(x) < 0$, se $x < 0$. Conclui-se portanto que, para qualquer $x_0 > 2/c$, todas as iteradas são negativas e por conseguinte tal sucessão não pode convergir para $\tilde{x} = 1/c$. Finalmente, se $x_0 \in (1/c, 2/c)$ o ponto x_1 pertence ao intervalo $(0, 1/c)$ e as iteradas subsequentes formam uma sucessão crescente convergente para $\tilde{x}$. Com efeito, escolhido um qualquer ponto inicial em $(0, 1/c)$, dado que a função g é estritamente crescente nesse intervalo (e limitada por $\tilde{x} = 1/c$), a correspondente sucessão converge monotonamente para o ponto fixo em causa.

3 (a) Sejam $f(x_1, x_2, x_3) = (x_1^2 - x_2 + x_3^2 - \alpha, -x_1 + x_2 - x_3 - \beta, 2\,x_1 - x_2 + 3\,x_3 - \gamma)^T$ e $x^{(0)} = (1, 0, 1)^T$. Como

$$J_f(x_1, x_2, x_3) = \begin{bmatrix} 2\,x_1 & -1 & 2\,x_3 \\ -1 & 1 & -1 \\ 2 & -1 & 3 \end{bmatrix},$$

o sistema $A\,u = b$ tem por matriz

$$A = J_f(x^{(0)}) = \begin{bmatrix} 2 & -1 & 2 \\ -1 & 1 & -1 \\ 2 & -1 & 3 \end{bmatrix}$$

e segundo membro

$$b = -f(x^{(0)}) = -(2-\alpha, -2-\beta, 5-\gamma)^T = (\alpha-2, 2+\beta, \gamma-5)^T.$$

Cálculo da matriz A^{-1}:

$$\left[\begin{array}{ccc|ccc} 2 & -1 & 2 & 1 & 0 & 0 \\ -1 & 1 & -1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 2 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 3 & 0 & -2 \\ 0 & 1/2 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & 4 & 2 & -2 \\ 0 & 1/2 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 3 & 0 & -2 \\ 0 & 1/2 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 2 & 1 & -1 \\ 0 & 1 & 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right] = [I|A^{-1}]\ .$$

Assim,

$$\begin{array}{l} ||A||_1 = max(5,3,6) = 6 \\ ||A^{-1}||_1 = max(4,3,2) = 4 \\ cond_1(A) = ||A||_1\,||A^{-1}||_1 = 24 < 25\ . \end{array}$$

3 (b) De $(2-\alpha, -2-\beta, 5-\gamma)^T = (1,0,1)^T$, tem-se $\alpha = 1, \beta = -2, \gamma = 4$.

3 (c) A matriz A é simétrica. Como os seus menores principais são $A_1 = 2 > 0, A_2 = 1 > 0$ e $A_3 = det(A) = 2 \times 2 + (-3+2) + 2\,(1-2) = 1 > 0$, a matriz é definida positiva. O método de Gauss-Seidel aplicado ao sistema converge para a sua solução, independentemente do vector inicial considerado, em particular quando $x^{(0)} = b$.

A.2.28

Exame de 30 de Junho 2016 (Parte 2)

1) Seja $a > 0$. Considere o sistema linear de equações $x_1 + a\,x_2 = 1+a$, $x_1 + 3\,x_2 + x_3 = 5$, $x_2 + x_3 - 2 = 0$.

(a) Escolha um valor de a, tal que $a \neq 1$, para o qual possa garantir convergência do método de Jacobi quando aplicado ao sistema anterior. Justifique. **[1.0]**

(b) Sendo $a = 1$, se tomar como aproximação inicial do sistema $x^{(0)} = (1,1,0)$, quantas iterações do método deverá efectuar para garantir um erro absoluto não superior a $2/3$, na norma $||\,||_\infty$? Justifique. **[1.0]**

2) Considere a tabela

x_i	0	-2	-4	1
y_i	1	γ	1	1

[1.0] **(a)** Suponha que $y_i = f(x_i)$, $i = 0, 1, 2, 3$, sendo f um polinómio de grau menor ou igual a dois. Calcule tal polinómio e obtenha de seguida o valor de γ. Apresente todos os cálculos que efectuar e justifique.

[1.0] **(b)** Na tabela dada substitua γ pelo valor 2. Calcule o coeficiente de x^3 do polinómio interpolador de Newton dos valores tabelados. Justifique.

[1.0] **(c)** Para $\gamma = 1$, escreva explicitamente a função a minimizar ao aproximar os valores tabelados por funções aproximantes do tipo, $k(x) = a + b\,x + c\,x^2 + d\,x^3,\ a, b, c, d \in \mathbb{R}$, mediante o critério de mínimos quadrados. Atendendo a que a tabela contém 4 pontos, qual é a soma dos quadrados dos desvios da melhor aproximação? Justifique.

[1.5] **3 a)** Obtenha uma aproximação de $\displaystyle\int_0^1 \sin(2\,x)\,dx$, com erro inferior a 0.005, aplicando a regra de Simpson. Justifique.

[1.5] **(b)** Dada uma qualquer função integrável f, para calcular uma aproximação de $\int_a^b f(x)\,dx$, considere a regra do ponto médio $M(f) = (b-a) \times f\left((a+b)/2\right)$. Diga, justificando, se é verdadeira ou falsa a afirmação "a regra tem grau de precisão um". Apresente todos os cálculos que efectuar.

4) Considere o problema de valores iniciais

$$\begin{cases} y''(t) - t\,y^2(t) + y(t) = 1 \\ y(1) = 1, \quad y'(1) = -1, \qquad t \in [1, 1.5] \end{cases}$$

[0.5] **(a)** Escreva o problema dado como um sistema de equações diferenciais de primeira ordem.

[1.5] **(b)** Adoptando o passo $h = 0.1$, obtenha uma aproximação de $y'(1.1)$ mediante aplicação do método do ponto médio. Apresente todos os cálculos que efectuar.

Resolução

1 (a) As fórmulas de iteração do método escrevem-se

$$\begin{cases} x_1^{(k+1)} &= 1 + a - a\,x_2^{(k)} \\ x_2^{(k+1)} &= \left(5 - (x_1^{(k)} + x_3^{(k)})\right)/3, \qquad\qquad k = 0, 1, \ldots\ . \\ x_3^{(k+1)} &= 2 - x_2^{(k)} \end{cases}$$

Logo, a matriz de iteração do método C_J é:

$$C_J = \begin{bmatrix} 0 & -a & 0 \\ -1/3 & 0 & -1/3 \\ 0 & -1 & 0 \end{bmatrix}.$$

Como $det(C_J - \lambda I) = -\lambda\,(\lambda^2 - 1/3) + a\,\lambda/3 = p_2(\lambda)$, tem-se que $p_2(\lambda) = 0$ se e só se

$$\lambda\,(\lambda^2 - (1+a)/3) = 0 \Leftrightarrow \lambda = 0 \quad \text{ou} \quad \lambda = \pm\sqrt{(1+a)/3}\ .$$

Dado que $a > 0$, a condição necessária e suficiente de convergência do método, $\rho(C_J) < 1$, equivale neste caso a $0 < a < 2$. Assim, por exemplo para $a = 1/2$ o método é convergente para a solução $x = (1, 1, 1)$ do sistema. Para $a = 2$ o método não converge (excepto se $x^{(0)} = x$.)

1 (b) Seja $a = 1$ e $x^{(0)} = (1, 1, 0)$. Como

$$x^{(1)} = (1, 4/3, 1) \Longrightarrow x - x^{(1)} = (0, -1/3, 0) \Longrightarrow ||x - x^{(1)}||_\infty = 1/3 < 2/3\ .$$

A condição de erro enunciada é satisfeita efectuando uma iteração.

2 a) O polinómio interpolador nos nós $\{0, -4, 1\}$ é, evidentemente, $p(x) = 1$. Por conseguinte, $\gamma = p(-2) = 1$.
De facto, recorrendo à fórmula de Lagrange, tem-se

$$p(x) = \frac{(x+4)\,(x-1)}{4 \times (-1)} + \frac{x\,(x-1)}{(-4) \times (-5)} + \frac{x\,(x+4)}{5},$$

polinómio que interpola a tabela nos nós considerados. Assim,

$$\begin{aligned} \gamma = p(-2) \ &= \frac{2 \times (-3)}{-4} + \frac{(-2) \times (-3)}{20} + \frac{(-2) \times 2}{5} \\ &= 3/2 + 3/10 - 4/5 = 1\ . \end{aligned}$$

2 b) Levando em consideração o polinómio $p(x)$ anteriormente calculado, o polinómio interpolador da tabela com 4 nós é

$$p_3(x) = p(x) + c\,x\,(x+4)\,(x-1),$$

onde c é o coeficiente de x^3. Assim,

$$\gamma = p_3(-2) = p(-2) + c\,(-2)\,(2)\,(-3) = 1 + 12\,c\ .$$

Por conseguinte, dado que $\gamma = 2$,

$$c = (\gamma - 1)/12 = 1/12\ .$$

O mesmo resultado poderá ser obtido efectuando as diferenças divididas

x_i	y_i	$y[..]$	$y[...]$	$y[....]$
0	1			
		$-1/2$		
-2	2		$-1/4$	
		$1/2$		$1/12$
-4	1		$-1/6$	
		0		
1	1			

Atendendo à forma do polinómio interpolador de Newton dos valores tabelados, o coeficiente c de x^3 desse polinómio é a última diferença dividida calculada, isto é, $c = 1/12$.

2 (c) A função a minimizar é $F(a,b,c,d) = \sum_{i=0}^{3}(k(x_i) - y_i)^2 = (a-1)^2 + (a - 2\,b + 4\,c + 8\,d - 1)^2 + (a - 4\,b + 16\,c - 64\,d - 1)^2 + (a + b + c + d - 1)^2$. Designando por $p_3(x)$ o polinómio interpolador da tabela, uma vez que para $k(x) = p_3(x)$, são satisfeitas as igualdades $k(x_i) = y_i$, para $i = 0, 1, 2, 3$, donde se conclui que a melhor aproximação de mínimos quadrados dos valores tabelados coincide com o polinómio interpolador (neste caso $k(x) = p_3(x) = 1$). Por conseguinte, a soma dos quadrados dos desvios é nula.

3 (a) O valor exacto do integral é

$$I(f) = \int_0^1 \sin(2\,x)dx = -1/2\,(\cos(2\,x))|_{x=0}^{x=1} = 0.7080734\cdots \ .$$

Para $N = 2$ subintervalos resulta

$$h = 1/2 \Longrightarrow S_N(f) = h/3\,[f(0) + f(1) + 4f(1/2)] \simeq 0.712530 \ .$$

Logo,

$$I(f) - S_N(f) = -0.0044\cdots \ .$$

Conclui-se que basta fazer $N = 2$ para garantir uma aproximação do integral com erro absoluto menor do que $\epsilon = 0.005$. Usando uma majoração do erro, visto que $f \in C^4([0,1]$, tem-se

$$f^{(4)}(x) = 16\,\sin(2\,x) \Longrightarrow |f^{(4)}(x)| \leq 16, \ \forall x \in [0,1] \ .$$

Assim, sendo N o número de subintervalos a determinar,

$$|E_N(f)| = |I(f) - S_N(f)| \leq 1/180 \times 16/N^4 < \epsilon \Leftrightarrow N > \left(\frac{4}{45\,\epsilon}\right)^{1/4} \simeq 2.05 \ .$$

O número par imediatamente superior a 2.05 é $N = 4$. Por conseguinte, a seguinte aproximação de $I(f)$ satisfaz o critério de erro pretendido:

$$h = 1/4 \Longrightarrow S_N(f) = h/3\,[f(0) + f(1) + 4\,(f(1/4) + f(3/4)) + 2\,f(1/2)] \simeq 0.708327 \ .$$

3 (b) Considere-se a base dos polinómios de grau menor ou igual a 1, constituída pelos elementos $\phi_0(x) = 1$ e $\phi_1(x) = (x - (a+b)/2)$. Dado que

$$\begin{array}{ll} M(\phi_0) = b - a, & I(\phi_0) = \int_a^b \, dx = b - a \\ M(\phi_1) = 0, & I(\phi_1) = \int_a^b \, (x - (a+b)/2) \, dx = 0, \end{array}$$

conclui-se que a regra em causa é de grau de precisão ≥ 1. Mas, por exemplo para o polinómio $q(x) = x^2$, tem-se

$$M(q) = (b-a) \times ((a+b)/2)^2, \qquad I(q) = \int_a^b x^2 dx = (b^3 - a^3)/3 \ .$$

Como $M(q) \neq I(q)$ a regra possui grau 1 de precisão.

4 (a) Como $y''(t) = 1 + t\,y^2(t) - y(t)$, denotando y_1, y_2, por $y_1(t) = y(t)$ e $y_2(t) = y'(t)$, resulta o sistema de primeira ordem

$$\begin{cases} y_1'(t) = y_2(t) \\ y_2'(t) = 1 + t\,y_1^2(t) - y_1(t), \end{cases}$$

com $y_1(1) = 1$ e $y_2(1) = -1$.

4 (b) Fazendo $Y = (y_1, y_2)$, o sistema anterior é da forma

$$Y' = F(t, Y) = (y_2, 1 + t\,y_1^2 - y_1) \ .$$

Denotando por $\tilde{Y}$ o ponto a calcular em cada passo do método do ponto médio, a partir de um ponto Y, tem-se

$$\tilde{Y} = Y + h\,F(t + h/2, Y + h/2\,F(t, Y)) = Y + h\,F(t + h/2, Z) \ . \qquad (*)$$

Ora,

$$\begin{aligned} &Z = Y + h/2\,F(t, Y) = (z_1, z_2) = \\ &= (y_1, y_2) + h/2\,(y_2, 1 + t\,y_1^2 - y_1) = \big(y_1 + h/2\,y_2, y_2 + h/2\,(1 + t\,y_1^2 - y_1)\big) \, . \end{aligned}$$

Assim,

$$\begin{aligned} &F(t + h/2, Z) = (z_2, 1 + (t + h/2)\,z_1^2 - z_1) = \\ &= \big(y_2 + h/2\,(1 + t\,y_1^2 - y_1), 1 + (t + h/2)\,(y_1 + h/2\,y_2)^2 - (y_1 + h/2\,y_2)\big) \ . \end{aligned}$$

Por conseguinte, as expressões do método correspondente a $(*)$ escrevem-se

$$\begin{aligned} y_{1,i+1} &= y_{1,i} + h\,\Big[y_{2,i} + h/2\,(1 + t_i\,y_{1,i}^2 - y_{1,i})\Big] \\ y_{2,i+1} &= y_{2,i} + h\,\Big[1 + (t_i + h/2)\,(y_{1,i} + h/2\,y_{2,i})^2 - (y_{1,i} + h/2\,y_{2,i})\Big] \, , \qquad i = 0, 1, \ldots , \end{aligned}$$

sendo $y_{1,0} = 1$, $y_{2,0} = -1$, $t_0 = 1$, $h = 0.1$ e $i = 0$, obtém-se

$$\begin{aligned} y_{2,1} &= y_{2,0} + h\,\big[1 + (1 + h/2)\,(y_{1,0} + h/2\,y_{2,0})^2 - (y_{1,0} + h/2\,y_{2,0})\big] \\ &= -0.9002375 \simeq y'(0.1) \ . \end{aligned}$$

A.2.29

Teste 5 de Abril 2017

1) Seja $f(\epsilon) = \ln(1 + \epsilon)$.

(a) Obtenha um valor aproximado de $z = f(1/k)$, onde k representa o seu número de aluno, efectuando os cálculos num sistema de ponto flutuante com 4 dígitos na mantissa e arredondamento simétrico. Calcule o erro relativo do seu resultado, expresso em percentagem. **[1.0]**

(b) A função usada na alínea anterior é bem condicionada para valores de ϵ próximos de 0 ? Justifique. **[1.0]**

2) Pretende-se aproximar o valor $z = \sqrt[3]{7}$ mediante um processo iterativo.

[1.0] **(a)** Escreva uma equação não linear cuja solução seja o número z considerado. Diga, justificando se poderá partir do intervalo $I = [0, 2]$ para aproximar z através do método da bissecção.

[1.5] **(b)** Mostre que a sucessão $x_{m+1} = x_m + 1/8\,(7 - x_m^3)$, $m = 0, 1, \ldots$, com $x_0 \in [1, 2]$, converge para z . A convergência é linear? Justifique.

[1.5] **(c)** Use a função que adoptou na alínea 2(a) para obter a fórmula iterativa do método de Newton. Escolha um intervalo onde possa garantir convergência do método para z, qualquer que seja o valor inicial que considere. Justifique.

[1.0] **(d)** Aplique a fórmula anterior, com $x_0 = 2$, e determine a iterada x_2 do método de Newton. Atendendo a que $z \in [1.9, 2]$, obtenha uma majoração do erro absoluto de x_2 .

3) Considere o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}, \qquad b = (4, 5, 5)^T \qquad (*)$$

Para aproximar a solução deste sistema pretende-se usar um método iterativo da forma $x^{(n+1)} = C(\beta)\,x^{(n)} + g(\beta)$, onde $\beta < 0$ e

$$C(\beta) = \begin{bmatrix} 1+\beta & \beta/2 & \beta/2 \\ \beta/3 & 1+\beta & \beta/3 \\ \beta/2 & \beta & 1+\beta \end{bmatrix}, \qquad g(\beta) = (-2\,\beta, -5\,\beta/3, -5\,\beta/2)^T \qquad (**)$$

Sabe-se que os valores próprios da matriz $C(\beta)$ são os seguintes: $\lambda_1 = 1 + 2\,\beta$; $\lambda_{2,3} = 1 + \beta/2$.

[1.0] **(a)** Mostre que para $\beta = -1$ o processo anterior coincide com o método de Jacobi.

[1.0] **(b)** Diga, justificando, se o método de Jacobi converge, quando aplicado ao sistema $(*)$.

[1.0] **(c)** Mostre que, com $\beta = -1/2$, o método $(**)$ converge *para a solução exacta* $x = (1, 1, 1)$.

Resolução

1 (a) Dado que $k > 10^4$, resulta que $\epsilon = 1/k < 10^{-4}$. Assim, $1 + \epsilon = 1.0000\cdots$. Por conseguinte, $fl(1 + \epsilon) = 1 = +0.1000 \times 10^1$. Donde,

$$\bar{z} = fl(\ln(fl(1 + \epsilon))) = fl(0) = 0 \ .$$

Logo,

$$|\delta_{\bar{z}}| = \left|\frac{z - \bar{z}}{z}\right| = \left|\frac{z}{z}\right| = 1 = 100\,\% \ .$$

1 (b) De $f(\epsilon) = \ln(1+\epsilon)$, obtém-se

$$\frac{\epsilon\, f'(\epsilon)}{f(\epsilon)} = \frac{\epsilon}{(1+\epsilon)\,\ln(1+\epsilon)},$$

donde,

$$\lim_{\epsilon\to 0} \frac{\epsilon\, f'(\epsilon)}{f(\epsilon)} = \lim_{\epsilon\to 0} \frac{1}{\ln(1+\epsilon)+1} = 1\ .$$

Atendendo a que $cond_f(\epsilon) = |\epsilon\, f'(\epsilon)/f(\epsilon)|$, tem-se

$$\lim_{\epsilon\to 0} cond_f(\epsilon) = 1,$$

pelo que a função é bem condicionada para valores de ϵ próximos de 0 .

2 (a) Para $f(x) = x^3 - 7, \quad I = [0,2]$. Uma vez que f é contínua e $f(1) = 1 - 7 = -6 < 0$, $f(2) = 8 - 7 = 1 > 0$, o método da bissecção pode ser utilizado e converge para a (única) raiz $z = \sqrt[3]{7}$ no intervalo considerado.

2 (b) Sejam

$$g(x) = x + \frac{1}{8}\,(7 - x^3), \quad I = [1,2], \quad \text{tal que} \quad g \in C^1(I) \text{ e } g'(x) = 1 - 3/8\, x^2\ .$$

Dado que

$$\begin{array}{l} g(1) = 7/4 = 1.75 \in I, \quad g(2) = 15/8 = 1.875 \in I, \\ g'(x) = 0 \quad \text{se e só se} \quad 3\,x^2/8 = 1 \ \Leftrightarrow x = \sqrt{8/3} \simeq 1.63, \end{array}$$

a função g é monótona crescente em $[1, \sqrt{8/3}]$ e monótona decrescente em $\sqrt{8/3}, 2]$. Além disso,

$$max_{x\in[1,2]}\, g(x) = g(\sqrt{8/3}) \simeq 1.96 < 2\ .$$

Assim,

$$g(I) \subset I\ .$$

A função g' é decrescente em I, com $g'(1) = 5/8$ e $g'(2) = -1/2$. Por conseguinte,

$$L = max_{x\in I}\, |g'(x)| = g'(1) = 5/8 < 1\ .$$

Uma vez que

$$0 < 1/2 \le |g'(x)| \le 5/8 \quad \forall x \in I \implies 0 < |g'(z)| < 1,$$

a convergência é linear.

2 (c) Sendo $f(x) = x^3 - 7$ e, por exemplo, $I = [1.6, 2]$, tem-se

$$\begin{array}{l} x_0 \in I \\ x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)} = \dfrac{1}{3}\left(2\,x_k + \dfrac{7}{x_k^2}\right), \quad k = 0, 1, \ldots \end{array}$$

Dado que

$$\begin{array}{l} f \in C^2(I) \\ f'(x) = 3\,x^2 > 0, \quad \forall x \in I \\ f''(x) = 6\,x > 0, \quad \forall x \in I \\ \left|\dfrac{f(1.6)}{f'(1.6)}\right| \simeq 0.38 < 0.4, \qquad \left|\dfrac{f(2)}{f'(2)}\right| \simeq 0.08 < 0.4, \end{array}$$

sabe-se que uma vez escolhido um valor inicial $x_0 \in I$, o método converge para a raiz (única), z, da equação $f(x) = 0$, e a convergência é quadrática.

2 (d) Sendo $f(x) = x^3 - 7, \quad I = [1.9, 2]$, e $x_0 = 2$,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = \frac{1}{3}\left(2\,x_k + \frac{7}{x_k^2}\right), \quad k = 0, 1, \ldots$$

Assim,

$$\begin{array}{l} x_1 = x_0 - f(x_0)/f'(x_0) = 23/12 \simeq 1.916667 \\ x_2 = x_1 - f(x_1)/f'(x_1) \simeq 1.9129385 \end{array}$$

$$x_3 = x_2 - f(x_2)/f'(x_2) \simeq 1.91293118\ .$$

Fazendo

$$\mathcal{K} = \frac{1}{2}\frac{max|f''(x)|}{min_{x\in I}|f'(x)|} = \frac{1}{2}\,\frac{12}{3 \times 1.9^2} \simeq 0.5540,$$

tem-se

$$|z - x_2| \quad \leq \mathcal{K}\left(\frac{1}{\mathcal{K}}\,|z - x_0|\right)^4\ .$$

Atendendo a que $x_0, z \in [1.9, 2]$, resulta $|z - x_0| < 0.1$, logo

$$|z - x_2| \leq \mathcal{K}\left(\frac{1}{\mathcal{K}}\,0.1\right)^4 \simeq 0.0006\ .$$

Uma estimativa mais realista do erro de x_2 pode ser facilmente obtida:

$$z - x_2 \simeq x_3 - x_2 \simeq -7.3 \times 10^{-6}\ .$$

3 (a) O método de Jacobi aplicado ao sistema dado é gerado pela função iteradora

$$J(x) = D^{-1}\,(L + U)\,x + D^{-1}\,b = \begin{bmatrix} 0 & -1/2 & -1/2 \\ -1/3 & 0 & -1/3 \\ -1/2 & -1 & 0 \end{bmatrix} x + \begin{bmatrix} 2 \\ 5/3 \\ 5/2 \end{bmatrix}\ .$$

Assim, quando $\beta = -1$, o método $x^{(n+1)} = C(-1)\,x^{(n)} + g(-1)$ coincide com o método de Jacobi.

3 (b) Uma vez que

$$\rho(C_J) = \rho\,(C(-1)) = max(1, 1/2, 1/2) = 1,$$

o método de Jacobi não converge para a solução do sistema dado.

3 (c) Atendendo a que

$$\rho\left(C(-1/2)\right) = max(0, 3/4, 3/4) = 3/4 < 1,$$

o método $x^{(n+1)} = C(-1/2)\, x^{(n)} + g(-1/2)$ é convergente. Para mostrar que converge para $x = (1, 1, 1)$, basta mostrar que este vector é ponto fixo da função iteradora, isto é, $x = C(-1/2)\, x + g(-1/2)$. Com efeito,

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/4 & -1/4 \\ -1/6 & 1/2 & -1/6 \\ -1/4 & -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 5/6 \\ 5/4 \end{bmatrix} .$$

A.2.30

(Teste 1 Junho 2017)

1) Aplicou-se o método de Newton ao sistema de equações algébricas não lineares **[1.0]**

$$\begin{cases} x_1 + \alpha\, x_2^3 = 0 \\ \alpha x_1^3 + x_2 = \beta \end{cases}$$

onde $\alpha, \beta \in \mathbb{R} \setminus \{0\}$; sendo $x^{(0)} = [0 \;\; 2]^T$, foi obtida a iterada $x^{(1)} = [8 \;\; 2]^T$. Com base nesta informação, determine α e β .

2) Considere os seguintes valores de uma função f:

x	0	1	4
$f(x)$	3	9	5

a) Usando todos os valores tabelados, determine uma aproximação de $f(3)$ pela fórmula interpoladora de Newton. **[1.0]**

b) Supondo que f é um polinómio de grau 3, da forma $f(x) = x^3/2 + a_2\, x^2 + a_1\, x + a_0$, determine o ponto do intervalo [0,4] para o qual o erro absoluto de interpolação quadrática é máximo. Qual é o valor do erro absoluto nesse ponto? **[1.5]**

c) Pretende-se obter a função do tipo $g(x) = c + d\cos(\pi\, x/2)$ $(c, d \in \mathbb{R})$ que melhor se ajusta aos valores tabelados no sentido dos mínimos quadrados. Escreva o respectivo sistema de equações normais (não é necessário resolvê-lo). **[1.5]**

d) Obtenha um valor aproximado do integral $\displaystyle\int_{-2}^{4} x\, f(x) dx$, usando uma regra de quadratura do tipo **[1.5]**

$$Q(f) = A_0 f(0) + A_1 f(1) + A_2 f(4) \; .$$

Comece por determinar os pesos de $Q(f)$ de modo a que o seu grau seja o mais alto possível.

e) Determine o grau da regra utilizada na alínea anterior. Justifique. **[1.0]**

3) Considere o problema de valores iniciais:

$$y'(x) = x + y^2(x), \qquad y(1) = 0 \ .$$

Para aproximar a solução deste problema, considere dois métodos numéricos, dados pelas fórmulas

$$(A) \quad y_{i+1} = y_i + h(x_i + y_i^2) + \frac{h^2}{2}\left(1 + 2y_i x_i + 2y_i^3\right)$$

$$(B) \quad y_{i+1} = y_i + \tfrac{h}{2}(x_i + y_i^2) + \frac{h}{2}x_{i+1} + \frac{h}{2}\left(y_i + h(x_i + y_i^2)\right)^2 \ .$$

[1.5] **a)** Identifique cada um dos métodos pelo seu nome, justificando, e diga qual a sua ordem.

[1.0] **b)** Obtenha dois valores aproximados de $y(1.4)$, efectuando:
i) um passo do método (A);
ii) um passo do método (B) .

Resolução

1) O sistema pode ser definido através das seguintes funções:

$$\begin{cases} f_1(x_1, x_2) = x_1 + \alpha\, x_2^3 \\ f_2(x_1, x_2) = \alpha\, x_1^3 + x_2 - \beta \end{cases} \qquad \text{donde} \quad J(x_1, x_2) = \begin{bmatrix} 1 & 3\alpha x_2^2 \\ 3\alpha x_1^2 & 1 \end{bmatrix} \ .$$

Para $x^{(0)} = [0 \ \ 2]$, obtém-se

$$J(0, 2) = \begin{bmatrix} 1 & 12\alpha \\ 0 & 1 \end{bmatrix}.$$

Como $f_1(0, 2) = 8\,\alpha \quad$ e $\quad f_2(0, 2) = 2 - \beta$, o sistema linear correspondente à primeira iteração do método de Newton tem a forma

$$\begin{bmatrix} 1 & 12\,\alpha \\ 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -8\,\alpha \\ \beta - 2 \end{bmatrix} \ .$$

Resolvendo o sistema obtém-se

$$w_2 = \beta - 2, \qquad w_1 = -8 - 12\,(\beta - 2) \ .$$

A primeira iterada do método de Newton é dada por

$$x^{(1)} = x^{(0)} + w = (0, 2) + (-8\,\alpha - 12\,(\beta - 2), \beta - 2) = (-8\,\alpha - 12\alpha\,(\beta - 2), \beta) \ .$$

Comparando com o valor conhecido de $x^{(1)}$, temos

$$\beta = 2, \qquad -8\,\alpha - 12\,\alpha\,(\beta - 2) = 8 \ .$$

Logo, $\beta = 2$ e $\alpha = -1$.

2(a) Comecemos por calcular as diferenças divididas de f:

$$f[0,1] = 9 - 3 = 6, \qquad f[1,4] = (5-9)/3 = -4/3,$$

$$f[0,1,4] = (f[1,4] - f[0,1])/4 = -11/6 \ .$$

Da fórmula interpoladora de Newton, obtém-se

$$P(x) = 3 + 6\,x - 11/6\,x\,(x-1) \ .$$

Assim, $P(3) = 3 + 18 - 11 = 10$.

2(b) Comecemos por observar que $f^{(3)}(x) = 3!/2 = 3$. Pela fórmula do erro de interpolação, temos

$$e_2(x) = \frac{f^{(3)}(\xi)}{3!} x(x-1)(x-4) = \frac{1}{2}\,x(x-1)(x-4), \quad \xi \in (0,4) \ .$$

Para determinar o ponto onde $|e_2|$ é máximo, é necessário calcular os zeros de $e_2'(x)$. Dado que $e_2'(x) = 1.5\ x^2 + 5\ x - 2$, os zeros deste polinómio são $x_{1,2} = \dfrac{5 \pm \sqrt{13}}{3}$. Logo, o valor máximo do erro absoluto de interpolação é $|e_2\,(2.87)| \simeq 3.033$.

2(c) Tendo em conta que as funções de base são $\phi_0(x) = 1, \quad \phi_1(x) = \cos(\pi\,x/2)$, resulta

$$\begin{array}{ll} (\phi_0,\phi_0) = \sum_{i=0}^{2} 1 = 3, & (\phi_0,\phi_1) = \sum_{i=0}^{2} \cos(\pi\,x_i/2) = 2, \\ (\phi_1,\phi_1) = \sum_{i=0}^{2} \cos(\pi\,x_i/2)^2 = 2, & (\phi_0,f) = \sum_{i=0}^{2} f_i = 17, \\ (\phi_1,f) = \sum_{i=0}^{2} \cos(\pi\,x_i/2) f_i = 8 \ . & \end{array}$$

Assim, o sistema normal tem a forma

$$\begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 17 \\ 8 \end{bmatrix} \ .$$

2(d) Para determinar os pesos A_0, A_1, A_2 usamos o método dos coeficientes indeterminados, o qual se traduz no seguinte sistema linear,

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 4 \\ 0 & 1 & 16 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \int_{-2}^{4} x dx \\ \int_{-2}^{4} x^2 dx \\ \int_{-2}^{4} x^3 dx \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \\ 60 \end{bmatrix},$$

de solução $A_0 = -9, A_1 = 12, A_2 = 3$. Por conseguinte, o valor aproximado do integral é

$$Q(f) = A_0 f(0) + A_1 f(1) + A_2 f(4) = -27 + 108 + 15 = 96 \ .$$

Nota: Para construir a regra de quadratura $Q(f)$ teve-se em consideração que a função a integrar é da forma $xf(x)$, o que faz com que as funções a integrar no segundo membro do sistema sejam x, x^2 e x^3 — e não $1, x, x^2$, como aconteceria se não existisse o factor x .

2(e) Por construção, a regra de quadratura tem, pelo menos, grau 2. Para determinar o grau de $Q(f)$, verifique-se se a regra é exacta quando f é um polinómio de grau 3 . Temos

$$Q(x^3) = 12 \times 1 + 3 \times 64 = 204 \quad \text{e} \quad I(x^3) = \int_{-2}^{4} x^4\, dx = 211.2\ .$$

Logo, $Q(f)$ tem grau dois de precisão.

3(a) O método (A) é o método de Taylor de segunda ordem. Para o justificarmos basta ter em conta que $f(x, y) = x + y^2$. Por conseguinte,

$$\frac{\partial f}{\partial x} = 1 \quad \text{e} \quad \frac{\partial f}{\partial y} = 2y\ .$$

Logo, $y''(x) = 1 + 2\,y\,(x + y^2) = 1 + 2\,y\,x + 2\,y^3$, donde resulta a forma apresentada.
O método (B) é o método de Heun (ou dos trapézios), o qual é de segunda ordem. Com efeito, temos

$$f(x, y) = x + y^2, \qquad f(x + h, y + h\,f(x, y)) = x + h + (y + h\,(x + y^2))^2.$$

Substituindo na fórmula do método de Heun, obtém-se imediatamente a expressão dada.

3(b) i) Método de Taylor de segunda ordem:

$h = 1.4 - 1 = 0.4,$
$y_1 = y_0 + h\,(x_0 + y_0^2) + \frac{h^2}{2}\left(1 + 2\,x_0 y_0 + 2\,y_0^3\right) = 0.4 + (0.4)^2/2 = 0.4 + 0.08 = 0.48\ .$

ii) Método de Heun: $y_1 = y_0 + \dfrac{h}{2} + \dfrac{h}{2}\,1.4 + \dfrac{h}{2}\,0.4^2 = 0 + 0.2 + 0.28 + 0.032 = 0.512\ .$

A.2.31

(Exame — 6 Julho 2017, Parte 1)

1) Pretende-se calcular a função $f(x) = \dfrac{e^x - 1}{x}$, com valores de x próximos de zero.

(a) Admitindo que $x = 10^{-8}$ e que efectua os cálculos numa máquina que utiliza o sistema de vírgula flutuante $VF(10, 6, -10, 10)$, com arredondamento simétrico, que resultado obtém? Qual o erro relativo desse valor? **[1.0]**

(b) Sabendo que a função exponencial satisfaz a igualdade aproximada $e^x \approx 1 + x + x^2/2$, quando $x \approx 0$, proponha uma fórmula alternativa para aproximar $f(x)$ quando x é próximo de 0. Qual a vantagem dessa fórmula em relação à que utilizou na alínea anterior? **[1.0]**

2) Pretende-se aproximar as raizes reais da função

$$f(x) = \frac{x^3}{3} - 2\ln x - 1 \ .$$

(a) Mostre que f tem exactamente duas raizes positivas e indique um intervalo, de comprimento inferior a 1, que contenha a menor delas (z_1 .) **[1.0]**

(b) Utilizando o método da bissecção, determine um intervalo de comprimento não superior a 0.2 que contenha z_2 (a maior raiz de f) . **[1.0]**

(c) Considere as funções

$$g(x) = \exp(a\,x^3 + b), \qquad h(x) = \sqrt[3]{3 + 6\ln(x)} \ .$$

(i) Determine a e b, de modo a que as raizes de f sejam os pontos fixos de g. **[0.5]**

(ii) Com base na teoria sobre o método do ponto fixo, mostre que a função g não pode ser utilizada como função iteradora para aproximar a maior raiz de f, enquanto a função h pode. **[1.2]**

(iii) Tomando como aproximação inicial $x_0 = 2$ e utilizando a função iteradora h, calcule x_2 . Obtenha um majorante do erro absoluto desta última aproximação. **[0.8]**

3) Considere o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} a & c & 0 & \dots & \dots & 0 \\ c & a & c & 0 & \dots & 0 \\ 0 & c & a & c & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & c & a & c \\ 0 & \dots & \dots & 0 & c & a \end{bmatrix}$$

(matriz quadrada $n \times n$), $b \in \mathbb{R}^n$, a e c são números reais, $a \neq 0$.

(a) Escreva as fórmulas iteradoras dos métodos de Jacobi e Gauss-Seidel, quando aplicados ao sistema $A\,x = b$. **[1.0]**

(b) Diga, justificando, que condição devem satisfazer a e c para que fique garantida a convergência do método de Jacobi, quando aplicado ao sistema $A\,x = b$. **[1.0]**

(c) Seja $n = 3$, $a = 3$, $c = 1$, $b = (4, 4, 4)$. Tomando $x^{(0)} = (1, 1, 1)$, efectue duas iterações do método de Jacobi. Obtenha um majorante de $\|x - x^{(2)}\|_\infty$, onde x é a solução exacta do sistema considerado. **[1.5]**

Resolução

1(a) Ao calcular $\exp(10^{-8})$ no sistema $VF(10, 6, -10, 10)$, devido ao arredondamento, obtém-se 1 . Logo, neste sistema, tem-se $\tilde{f}(10^{-8}) = (1-1)/10^{-8} = 0$. É fácil verificar

que $f(x) \approx 1$, quando x é próximo de 0 (o limite de f, quando x tende para 0, é um caso notável). Logo, o erro relativo do resultado obtido é

$$\frac{1-0}{1} = 1 \ .$$

1(b) Atendendo a que $\exp x \approx 1 + x + x^2/2$, temos

$$f(x) = \frac{\exp(x) - 1}{x} \approx 1 + \frac{x}{2} \ .$$

Assim, se usarmos esta fórmula com um valor de x próximo de 0 (por exemplo, $x = 10^{-8}$) obtém-se $\tilde{f}(10^{-8}) = 1$, o que é um resultado preciso. Isto acontece porque nesta última fórmula não ocorre cancelamento subtractivo, como acontecia na alínea a).

2(a) A função f considerada é continuamente diferenciável em $\mathbb{R}^+$. Temos

$$f'(x) = -\frac{2}{x} + x^2 \ .$$

Daqui resulta que f tem um único ponto de mínimo $\hat{x}$ que satisfaz a equação $\hat{x}^2 = 2/\hat{x}$, ou seja, $\hat{x} = \sqrt[3]{2}$. Logo, f é decrescente em $]0, \sqrt[3]{2}]$ e crescente em $[\sqrt[3]{2}, \infty[$.
Por outro lado, temos que $\lim_{x\to 0+} f(x) = +\infty$, e $f(\sqrt[3]{2}) = -0.795$ logo, pelo teorema de Bolzano, f anula-se pelo menos uma vez neste no intervalo $]0, \sqrt[3]{2}]$.
Do mesmo modo, uma vez que $\lim_{x\to\infty} f(x) = +\infty$, conclui- que no intervalo $[\sqrt[3]{2}, \infty[$ a função também se anula, pelo menos, uma vez. Dado que f' tem uma única raiz, conclui-se que f não pode ter mais que 2 raizes.
Finalmente, para indicar um intervalo de comprimento inferior a 1 que contenha z_1, basta verificar, por exemplo, que $f(0.5) > 0$ e $f(1) < 0$, pelo que $z_1 \in [0.5, 1]$.

2(b) Como sabemos, a maior raiz z_2 encontra-se no intervalo $[\sqrt[3]{2}, \infty[$. Para iniciar o método da bissecção consideremos, por exemplo, o intervalo $[a, b] = [1.5, 2]$.
Dado que $f(1.5) = -0.685 < 0$ e $f(2) = 0.280 > 0$, pelo que este intervalo contém z_2 .

No primeiro passo do método definimos

$$x_1 = (a + b)/2 = 1.75 \ .$$

Uma vez que $f(x_1) = -0.33 < 0$, concluímos que $z_2 \in [1.75, 2]$.
No segundo passo do método definimos

$$x_2 = (1.75 + 2)/2 = 1.875 \ .$$

Dado que $f(x_2) = -0.599 < 0$, conclui-se que $z_2 \in [1.875, 2]$. Este último intervalo tem comprimento inferior a 0.2 .

2(c-i) As raizes de f satisfazem

$$f(z) = \frac{z^3}{3} - 2\,\ln z - 1 = 0 \ .$$

Por conseguinte,

$$\frac{z^3}{3} - 1 = 2\,\ln z \quad \Leftrightarrow \quad \ln z = \frac{1}{2} - \frac{z^3}{6} \quad \Leftrightarrow \quad z = \exp(\frac{1}{2} - \frac{z^3}{6}) \ .$$

A última igualdade significa que z é um ponto fixo da função g indicada, onde $a = -1/6$ e $b = 1/2$.

2(c-ii) Comecemos por analisar a função g. Considerando $a = -1/6$ e $b = 1/2$, os pontos fixos desta função coincidem com as raizes de f . Em particular, esta função tem um ponto fixo z_2 no intervalo $[1.875, 2]$ (ver alínea b). Para esclarecer se esta função pode ou não ser utilizada como função iteradora, importa verificar como se comporta a sua primeira derivada no intervalo dado. Temos

$$g'(x) = x^2/2 \exp(-1/2 - x^3/6) \ .$$

Logo, $g'(1.875) = 3.199$ e $g'(2) = 4.602$. Além disso,

$$g''(x) = x \exp(-1/2 - x^3/6) + x^4/4 \exp(-1/2 - x^3/6) \ .$$

Assim, $g''(x) > 0$ e $g'(x)$ é crescente em $[1.875, 2]$. Em conclusão, temos $|g'(x)| > 1, \forall x \in [1.875, 2]$, pelo que g não pode ser utilizada para aproximar z_2 .

Analisemos agora a função h. Facilmente se verifica que os pontos fixos desta função também são raizes de f, e

$$h'(x) = \frac{2}{(3 + 6 \ln x)^{2/3} x} \ .$$

Pode verificar-se que h' é positiva e decrescente em $[1.875, 2]$. Além disso, $h'(1.875) = 0.298$ e $h'(2) = 0.269$, pelo que $L = \max_{x \in [1.875,2]} |h'(x)| < 1$.
Finalmente, temos $h(1.875) = 1.892$ e $h(2) = 1.927$, pelo que no intervalo $[1.875, 2]$ a função h satisfaz todas as condições do teorema do ponto fixo. Logo, a função h pode ser utilizada como função iteradora.

2(c-iii) Sendo $x_0 = 2$, temos $x_1 = 1.9273$ e $x_2 = 1.9073$. Da alínea anterior resulta,

$$L = \max_{x \in [1.875,2]} |h'(x)| = 0.298 \ .$$

Aplicando a estimativa do erro, temos

$$|x_2 - z_2| \leq \frac{L}{1 - L} |x_2 - x_1| = 0.00855 \ .$$

3(a) Tendo em conta que cada linha da matriz tem, no máximo, três entradas diferentes de zero, as fórmulas iteradoras do método de Jacobi escrevem-se, para $k = 0, 1, \ldots,$

$$\begin{aligned}
x_1^{(k+1)} &= \frac{b_1 - c\, x_2^{(k)}}{a} \\
&\vdots \\
x_i^{(k+1)} &= \frac{b_i - c\, x_{i+1}^{(k)} - c\, x_{i-1}^{(k)}}{a}, \qquad i = 2, ..., n-1 \\
&\vdots \\
x_n^{(k+1)} &= \frac{b_n - c\, x_{n-1}^{(k)}}{a} \ .
\end{aligned}$$

No caso do método de Gauss-Seidel, temos

$$\begin{aligned}
&x_1^{(k+1)} = \tfrac{b_1 - c\,x_2^{(k)}}{a} \\
&\vdots \\
&x_i^{(k+1)} = \tfrac{b_i - c\,x_{i+1}^{(k)} - c x_{i-1}^{(k+1)}}{a}, \qquad i = 2, ..., n-1 \\
&\vdots \\
&x_n^{(k+1)} = \tfrac{b_n - c\,x_{n-1}^{(k+1)}}{a} \ .
\end{aligned}$$

3(b) Para que fique garantida a convergência do método de Jacobi (ou do de Gauss-Seidel), basta que a matriz A tenha a diagonal estritamente dominante por linhas, ou seja, o elemento da diagonal principal de cada linha deve ser, em valor absoluto, superior à soma módulos dos outros elementos da mesma linha. Neste caso, dada a forma da matriz a condição é

$$|a| > 2\,|c| \ .$$

3(c) Para os valores dados de n, b, c e a as fórmulas iteradoras do método de Jacobi são, para $k = 0, 1, \ldots$,

$$x_1^{(k+1)} = \frac{4 - x_2^{(k)}}{3}$$

$$x_2^{(k+1)} = \frac{4 - x_3^{(k)} - x_1^{(k)}}{3}$$

$$x_3^{(k+1)} = \frac{4 - x_2^{(k)}}{3} \ .$$

Primeira iteração:

$$x_1^{(1)} = \frac{4-1}{3} = 1$$

$$x_2^{(1)} = \frac{4-1-1}{3} = \frac{2}{3}$$

$$x_3^{(1)} = \frac{4-1}{3} = 1 \ .$$

Segunda iteração:

$$x_1^{(2)} = \tfrac{4-2/3}{3} = \frac{10}{9}$$

$$x_2^{(2)} = \tfrac{4-1-1}{3} = \frac{2}{3}$$

$$x_3^{(3)} = \tfrac{4-2/3}{3} = \frac{10}{9} \ .$$

Para obter a estimativa do erro, é necessário determinar a norma da seguinte matriz de iteração C,

$$C = -D^{-1}(L+U) = \begin{bmatrix} 0 & -1/3 & 0 \\ -1/3 & 0 & -1/3 \\ 0 & -1/3 & 0 \end{bmatrix} .$$

Logo, $\|C\|_\infty = 2/3$. Por outro lado, $\|x^{(2)} - x^{(1)}\|_\infty = 10/9 - 1 = 1/9$. Por conseguinte,

$$\|x - x^{(2)}\|_\infty \leq \frac{\|C\|_\infty}{1 - \|C\|_\infty} \|x^{(2)} - x^{(1)}\|_\infty = 2/9 \ .$$

A.2.32

(Exame — 6 Julho 2017, Parte 2)

1) Considere funções da forma $\phi(x) = a^k\, p(x) + b^k\, q(x),\ x \in [-1, 1]$, onde k é um inteiro e $a, b \in \mathbb{R}$. São conhecidos os valores da seguinte tabela:

x_i	-1	0
$p(x_i)$	1	0
$q(x_i)$	0	1

(a) Para $k = 2$, escreva o sistema não linear (de incógnitas a e b), satisfazendo as condições $\phi(-1) = 4$ e $\phi(0) = 9$. Dada uma aproximação inicial, diferente de $(0, 0)$, obtenha explicitamente as fórmulas do método de Newton aplicado ao sistema. **[1.0]**

(b) Sendo $Z = (z_1, z_2)$, com $z_1, z_2 > 0$, solução exacta do sistema anterior e Z_n a n-ésima iterada do método de Newton, efectuando cálculos exactos mostre que se usar para aproximação inicial $Z_0 = (1, 1)$, se tem $||Z - Z_2||_\infty = 2/5$. **[1.0]**

2) Levando em conta os valores de p e q dados na tabela em **1**), e sabendo que $p(1) = q(1) = 0$, considere a função $g(x) = p(x) + q(x)$, onde $p, q \in C^\infty([-1, 1])$.

(a) Usando a base de Lagrange associada aos nós $\{-1, 0, 1\}$, obtenha a expressão do polinómio interpolador de g nesses nós. Justifique. **[1.5]**

(b) Admitindo que, para i=0,1,..., são satisfeitas as desigualdades $|p^{(i)}(x)| \leq 1/2$ e $|q^{(i)}(x)| \leq 1/3$, calcule um majorante do erro absoluto de interpolação no ponto $x = 1/2$. Justifique. **[1.0]**

3 a) Seja $I(f) = \displaystyle\int_{1.3}^{1.5} e^{-2\,x}\, dx$. Obtenha uma aproximação de $I(f)$ mediante aplicação da regra de Simpson simples. **[1.0]**

(b) Subdividindo o intervalo de integração em 4 partes, diga qual o erro que deverá resultar da regra de Simpson composta quando aplicada à função $g(x) = 1 + \alpha\,(x - 1.3) + \beta\,(x - 1.3)(x - 1.35)$, sendo α, β constantes arbitrárias. Justifique. **[1.0]**

4) Considere o problema de valores iniciais

$$\begin{cases} y'(t) - p(t) - q(t) = 0 \\ y(-1) = 1, \qquad\qquad\qquad\qquad t \in [-1, 1] \end{cases}$$

onde p e q são polinómios que interpolam os pontos da tabela dada em **1**) .

[0.5] **(a)** Mostre que $p(t) + q(t) = 1, \ \forall t\, [-1, 1]$.

[1.0] **(b)** Dado o passo $h > 0$, escreva a equação às diferenças do método de Taylor de segunda ordem aplicado ao problema anterior.

[1.0] **(c)** Para $h = 1$, efectue dois passos do referido método.

[1.0] **(d)** Sem calcular a solução do problema de valor inicial dado, mostre que $y(1)$ coincide com o valor que obteve na alínea anterior. Justifique.

Resolução

1 (a) O sistema a resolver tem a forma $a^2 = 4$ e $b^2 = 9$. Assim, $(a, b) = (\pm 2, \pm 3)$ são as suas soluções.
Fazendo $F(a, b) = (a^2 - 4, b^2 - 9)$, obtém-se

$$J(a,b) = \begin{bmatrix} 2\,a & 0 \\ 0 & 2\,b \end{bmatrix} \quad \text{e} \quad J^{-1}(a,b) = \begin{bmatrix} 1/(2\,a) & 0 \\ 0 & 1/(2\,b) \end{bmatrix}, \quad \text{onde} \quad a, b \neq 0 \ .$$

Por conseguinte, as fórmulas de iteração do método são

$$a_{k+1} = a_k - \frac{a_k^2 - 4}{2\,a_k} \quad = \frac{a_k^2 + 4}{2\,a_k}$$

$$b_{k+1} = b_k - \frac{b_k^2 - 9}{2\,b_k} \quad = \frac{b_k^2 + 9}{2\,b_k}, \qquad k = 0, 1, \ldots$$

(b)

$$Z_0 = (1, 1), \quad Z = (2, 3)$$

$$Z_1 = (5/2, 5), \qquad Z_2 = (41/20, 34/10) = (41/20, 17/5) \ .$$

Logo,

$$||Z - Z_2||_\infty = ||(-1/20, -2/5)||_\infty = 2/5 = 0.4 \ .$$

2 (a) Seja $P(x)$ o polinómio interpolador de g nos nós $\{-1, 0, 1\}$. Tem-se

$$P(x) = g(-1)\, l_o(x) + g(0)\, l_1(x) + g(1)\, l_2(x),$$

onde $l_0(x), l_1(x)$ e $l_2(x)$ são os elementos da base de Lagrange associado aos nós dados. Como

$$\begin{array}{ll} g(-1) & = p(-1) + q(-1) = 1 \\ g(0) & = p(0) + q(0) = 1 \\ g(1) & = p(1) + q(1) = 0 \end{array}$$

e

$$l_0(x) = \frac{x\,(x-1)}{(-1)\,(-2)} = \frac{x\,(x-1)}{2}$$

$$l_1(x) = \frac{(x+1)\,(x-1)}{-1} = -(x+1)\,(x-1)$$

resulta

$$P(x) = l_0(x) + l_1(x) = \frac{x\,(x-1)}{2} - (x+1)\,(x-1)\ .$$

2 (b) Dado que $g \in C^\infty$, é aplicável a fórmula de erro de interpolação. Assim,

$$|g(1/2) - P(1/2)| \leq \frac{g^{(3)}(\xi)}{6}\,|(1/2+1)\,1/2\,(1/2-1)|, \qquad \xi \in (-1,1)$$

Ora, de $g(x) = p(x) + q(x)$ resulta

$$\begin{array}{l} g^{(3)}(x) = p^{(3)}(x) + q^{(3)}(x), \quad \text{donde} \\ |g^{(3)}(x)| \leq 1/2 + 1/3 = 5/6 \quad \forall x \in [-1,1]\ . \end{array}$$

Por conseguinte,

$$|g(1/2) - P(1/2)| \leq 5/36\ (3/2 * 1/2 * 1/2) = 5/96 \simeq 0.052\ .$$

3(a) Para $h = 0.2/2 = 0.1$, obtém-se

$$S(f) = h/3\ [f(1.3) + 4f(1.4) + f(1.5)] = 0.1/3\left[e^{-2.6} + 4\,e^{-2.8} + e^{-3}\right] \simeq 0.012243363\ .$$

3(b) O erro deverá ser nulo porquanto a regra é exacta para qualquer polinómio de grau menor ou igual a três.

4(a) O polinómio $p(x) = -x$ interpola os pontos $((-1,1),(0,0))$ e o polinómio $q(x) = x + 1$ interpola $((-1,0),(0,1))$. Assim, $p(x) + q(x) = 1, \quad x \in [-1,1]$.

4(b) problema de valor inicial a resolver é da forma

$$\begin{array}{l} y'(t) = 1 \\ y(-1) = 1, \qquad -1 \leq t \leq 1\ . \end{array}$$

Neste caso o método de Taylor de qualquer ordem coincide com o método de Euler,

$$y_{i+1} = y_i + h \qquad i = 0, 1, \ldots$$

4(c) Para $h = 1$, $y_0 = 1$, resulta

$$\begin{array}{l} y_1 = 1 + 1 = 2 \\ y_2 = y_1 + 1 = 3\ . \end{array}$$

4(d) O erro de y_2 é nulo, isto é, $y_2 = y(1) = 3$. Com efeito, dado que $f(t, y) = 1$ e

$y(t + h) = y(t) + h\, y'(t) + h^2 y''(t) + \ldots$

obtém-se,

$$y(t + h) = y(t) + h \quad \Longrightarrow \quad y(t_{i+1}) = y(t_i) + h, \quad i = 0, 1, \ldots \quad \text{onde} \quad y(t_0) = y_0 = 1 \ .$$

Uma vez que no método de Euler $y_{i+1} = y_i + h, \quad i = 0, 1, \ldots$, conclui-se que o método é exacto para o p.v.i. dado, em particular para y_2. Com efeito, $y(1) = y(-1) + \int_{-1}^{1} dt = 1 + 2 = 3 \equiv y_2$.

A.2.33

(Teste — 30 Maio 2018)

1) Considere o sistema de equações não lineares

$$\begin{cases} x^2 + y^2 = 4 \\ x - 2\, y = 0 \end{cases}$$

Designe por $X^{(k)} = (x^{(k)}, y^{(k)})$ a k-ésima iterada do método de Newton aplicado ao sistema anterior.

[0.5] **(a)** Poderá iniciar o referido método com $X^{(0)} = (-1, 2)$? Justifique.

[1.0] **(b)** Partindo de $X^{(0)} = (2, 2)$, diga se $||X^{(1)} - X^{(0)}||_\infty < 2$. Justifique.

2) Dada a função $f(x) = -4\, x^3 + 2\, x - 2$, considere a tabela de pontos $\{x_i, f(x_i)\}$, sendo $x_i = -3 + i\, h$, com $h = 1$, para $i = 0, 1, 2, 3$.

[1.0] **(a)** Efectuando cálculos exactos, obtenha o polinómio interpolador de Newton associado aos <u>três primeiros</u> pontos da tabela.

[1.5] **(b)** Sem substituir f pela sua expressão, mostre que é válida a igualdade a seguir, onde $q(x)$ é o polinómio interpolador que calculou na alínea anterior :

$$f(x) + 4\,(x + 3)\,(x + 2)\,(x + 1) = q(x), \quad \forall x \in [-3, -1] \ .$$

[1.5] **(c)** Usando funções aproximantes do tipo $h(x) = \beta\, x + \alpha$, com $\alpha, \beta \in \mathbb{R}$, calcule a melhor aproximação de mínimos quadrados dos <u>três últimos</u> pontos tabelados (deverá efectuar cálculos exactos).

3) Para $f(x) = -4\, x^3 + 2\, x - 2$, seja $I(f) = \int_{-2}^{2} f(x)\, dx$.

[1.5] **(a)** Obtenha uma aproximação de $I(f)$, aplicando a regra dos trapézios composta, com 4 subintervalos. Indique qual o valor do passo de quadratura bem como a expressão que utilizou nos cálculos que efectuar.

[0.5] **(b)** Qual é o grau de precisão da regra anterior? Justifique.

[1.0] **(c)** Desprezando erros de arredondamento, se aplicasse a regra de Simpson composta, com 15 nós, qual o respectivo erro de quadratura? Justifique.

[1.5] **4)** Dado o problema de valor inicial $y'(t) = t^2, \quad y(-1) = 2$, aproxime $y(0)$ mediante aplicação do método de Euler, com passo $h = 1$. Calcule o erro do valor que obteve. Justifique.

Resolução

1(a) Sejam

$$f(x,y) = (x^2 + y^2 - 4, x - 2\,y)^T \quad \text{e} \quad J_f(x,y) = \begin{bmatrix} 2\,x & 2\,y \\ 1 & -2 \end{bmatrix} .$$

As sucessivas iteradas do método poderão ser calculadas desde que a matriz $J(X^k)$ seja não singular, isto é, $det(J_f(X^k)) \neq 0$. Ora,

$$det\,(J(x,y)) = 4\,x - 2\,y = -2\,(2\,x + y) .$$

Assim, para $X^{(0)} = (-1, 2)$, a matriz $J_f(X^{(0)})$ é singular pelo que não podemos aplicar o método iniciando-o em $X^{(0)}$.

1(b) Seja $\Delta X^{(0)} = X^{(1)} - X^{(0)}$. De

$$\begin{bmatrix} 4 & 4 \\ 1 & -2 \end{bmatrix} \Delta X^{(0)} = -f(X^{(0)}) = -\begin{bmatrix} 4 \\ -2 \end{bmatrix} = \begin{bmatrix} -4 \\ 2 \end{bmatrix}$$

obtém-se,

$$\Delta X^{(0)} = -\frac{1}{12}\begin{bmatrix} -2 & -4 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} -4 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Assim,

$$||\Delta X^{(0)}||_\infty = ||X^{(1)} - X^{(0)}||_\infty = 1 < 2\,.$$

2(a)

x_i	f_i	$f[..]$	$f[...]$
-3	100		
		-74	
-2	26		24
		-26	
-1	0		

Assim,

$$q(x) = 100 - 74\,(x+3) + 24\,(x+3)\,(x+2) .$$

2(b) A função dada é polinómio de grau 3, tal que $f^{(3)}(x) = -4 \cdot 3!$. Sabemos que existe $\xi \in (-3, -1)$, tal que

$$f(x) - q(x) = \frac{f^{(3)}(\xi)}{3!}\,(x+3)\,(x+2)\,(x-1) = -4\,(x+3)\,(x+2)\,(x-1), \quad \forall x \in [-3, -1],$$

donde a validade da igualdade em causa.

2(c) Para $x_0 = -2, x_1 = -1, x_2 = 0$, tem-se $f(x_0) = 26, f(x_1) = 0$ e $f(x_2) = -2$. Pretende-se minimizar

$$Q(\alpha,\beta) = \sum_{i=0}^{i=2} (\beta\, x_i + \alpha - f(x_i))^2, \quad \text{donde}$$

$$\begin{cases} \sum_{i=0}^{i=2} (\beta\, x_i + \alpha - f(x_i))\, x_i = 0 \\ \\ \sum_{i=0}^{i=2} (\beta\, x_i + \alpha - f(x_i)) = 0, \end{cases}$$

isto é,

$$\begin{array}{ll} \left(\sum_{i=0}^{i=2} x_i^2\right)\, \beta + \left(\sum_{i=0}^{i=2} x_i\right)\, \alpha & = \sum_{i=0}^{i=2} x_i\, f(x_i) \\ \left(\sum_{i=0}^{i=2} x_i\right)\, \beta + \left(\sum_{i=0}^{i=2} 1\right)\, \alpha & = \sum_{i=0}^{i=2} f(x_i)\ . \end{array}$$

Substituindo pelos valores de $x_i, f(x_i)$ considerados, resulta

$$\begin{cases} 5\,\beta & -3\,\alpha & = -52 \\ -3\,\beta & +3\,\alpha & = 24\ . \end{cases}$$

Assim,

$$\begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} -52 \\ 24 \end{bmatrix} = \begin{bmatrix} -14 \\ -6 \end{bmatrix},$$

pelo que a melhor aproximação pretendida tem a forma $h(x) = -14\,x - 6$.

3(a) Para $h = (b-a)/4 = 1$, $x_0 = -2$, $f(-2) = 26$, $x_1 = -1$, $f(-1) = 0$, $x_2 = 0$, $f(0) = -2$, $x_3 = 1$, $f(1) = -4$, $x_4 = 2$, $f(2) = -30$, obtém-se,

$$\begin{aligned} T_4(f) &= \frac{h}{2}\, [f(-2) + f(2) + 2\,(f(-1) + f(0) + f(1))] \\ \\ &= \frac{1}{2}\, [26 - 30 + 2\,(0 - 2 - 4)] = -8\ . \end{aligned}$$

3(b) A regra é de grau um, no sentido de que é exacta para qualquer polinómio de grau não superior a 1, mas já não é exacta para polinómios de grau ≥ 2, conforme se pode concluir da respectiva fórmula de erro de quadratura.

3(c) Dado que f é polinómio do terceiro grau e a regra de Simpson é de grau de precisão 3, esta regra é exacta para a função considerada. Assim, ao aplicar-se a regra com 15 nós, o respectivo erro será nulo (desprezando erros de arredondamento).

4) Como $y'(t) = t^2$, resulta $y(0) = y(-1) + \displaystyle\int_{-1}^{0} t^2\, dt = 2 + \frac{t^3}{3}|_{t=-1}^{t=0} = \frac{7}{3}$. Pelo método de Euler, para $t_0 = -1, y_0 = 2, h = 1$, obtém-se,

$$y_1 = y_0 + h\, f(t_0, y_0) = y_0 + h\, t_0^2 = 2 + (-1)^2 = 3\ .$$

Assim, $y(0) - y_1 = 7/3 - 3 = -2/3$.

A.2.34 (Exame – 5 Julho 2018, Parte 2)

1) Considere a função $G : \mathbb{R}^3 \setminus \{(0,0,0)\} \mapsto \mathbb{R}^3$, tal que $G(x,y,z) = (\ln(x^2) + \sin(y) - 1, \ln(y) + \sin(z^2), z - 1/2)$. Tomando para valores iniciais $x^{(0)} = 1$, $y^{(0)} = \pi/2$ e $z^{(0)} = \sqrt{\pi}$, pretende-se aplicar o método de Newton ao sistema não linear $G(x,y,z) = (0,0,0)$.

(a) Seja $v = (x^{(1)} - x^{(0)}, y^{(1)} - y^{(0)}, z^{(1)} - z^{(0)})^T$. Determine as matrizes M e c do sistema linear da forma $M\,v = c$ que lhe permite determinar a primeira iterada do método. Justifique. (Não é necessário resolver o sistema). [1.0]

(b) Sabendo que $||M^{-1}||_1 < 7$, mostre que $||v||_1 < 7\,(\sqrt{\pi} - 1/2 + \ln(\pi/2))$. [1.5]

2) Considere uma tabela de valores $\gamma = \{x_i, \phi(x_i)\}$, tal que $x_i = 1 + i\,h$, para $i = 1,2,3,4$, onde $h = 0.2$ e $\phi(x) = 1/\cos(2\,x)$.

(a) Sendo $A\,u = b$ o sistema de equações normais a partir do qual se pode calcular a melhor aproximação polinomial quadrática da tabela γ (no sentido dos mínimos quadrados), obtenha a matriz A, apresentando os resultados arredondados simetricamente para três casas decimais. Justifique. [1.0]

(b) Usando os 2 últimos nós da tabela γ, aplique o método dos coeficientes indeterminados para obter uma aproximação do integral $\int_{1.6}^{1.8} 1/\cos(2\,x)\,dx$. Qual é a designação habitual da regra de quadratura que construiu? Justifique. [1.0]

(c) Sabe-se que para $0 \leq j \leq 3$, $|\phi^{(j)}(x)| \leq 6$, $\forall x \in \mathbb{R}$. Se p designar o polinómio interpolador de Lagrange nos 3 primeiros nós da tabela, diga se o erro (absoluto) de interpolação, em $x = 1.35$ é menor do que 0.002 . Justifique. [1.0]

(d) Sendo p o polinómio interpolador referido em **(c)**, obtenha o valor do integral $\int_{1.2}^{1.6} p(x)\,dx$, sem calcular p . Justifique. Apresente o resultado arredondado (por corte) para três casas decimais. [1.5]

3) Considere o problema de valor inicial $y'(t) = t^2 - \cos(y(t))$, com
$$y(1) = 2, \quad t \in [1,2] .$$

(a) Escreva a fórmula de recorrência para o método de Heun, com passo $h > 0$. [1.0]

(b) Para $h = 0.2$, calcule uma aproximação de $y(1.2)$ mediante aplicação da fórmula anterior. [1.0]

(c) "O referido método, aplicado ao problema dado, tem segunda ordem de convergência no intervalo $[1,2]$". Que significado atribui a essa afirmação? Justifique. [1.0]

Resolução

1(a)

$$J_G(x,y,z) = \begin{bmatrix} 2/x & \cos(y) & 0 \\ 0 & 1/y & 2\,z\,\cos(z^2) \\ 0 & 0 & 1 \end{bmatrix} .$$

Para $X^{(0)} = (1, \pi/2, \sqrt{\pi})^T$, resulta

$$M = J_G(X^0) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2/\pi & -2\sqrt{\pi} \\ 0 & 0 & 1 \end{bmatrix}$$

e

$$c = -G(X^{(0)}) = (0, -\ln(\pi/2), 1/2 - \sqrt{\pi})^T \ .$$

1(b) Dado que $||c||_1 = \sqrt{\pi} - 1/2 + \ln(\pi/2)$, de $M\,v = c$, obtém-se

$$||v||_1 \leq ||M^{-1}||_1 \, ||c||_1 < 7\left(\sqrt{\pi} - 1/2 + \ln(\pi/2)\right) \ .$$

2(a) Sendo $f(x) = 1/\cos(2\,x)$, nós $x_0 = 1.2$, $x_1 = 1.4$, $x_2 = 1.6$, $x_3 = 1.8$, para as 'observações' $f = (f(x_0), f(x_1), f(x_2), f(x_3))^T$, e funções aproximantes polinomiais

$$\begin{array}{rl} p(x) & = a_0 + a_1\,x + a_2\,x^2, \quad a_0, a_1, a_2 \in \mathbb{R} \\ & = a_0\,\phi_0(x) + a_1\,\phi_1(x) + a_2\,\phi_2(x), \end{array}$$

sejam

$$\begin{array}{l} \phi_0 = (\phi_0(x_0), \phi_0(x_1), \phi_0(x_2), \phi_0(x_3))^T = (1,1,1,1)^T \\ \phi_1 = (\phi_1(x_0), \phi_1(x_1), \phi_1(x_2), \phi_1(x_3))^T = (x_0, x_1, x_2, x_3)^T \\ \phi_2 = (\phi_2(x_0), \phi_2(x_1), \phi_2(x_2), \phi_2(x_3))^T = (x_0^2, x_1^2, x_2^2, x_3^2)^T \ . \end{array}$$

A matriz A é da forma

$$\begin{bmatrix} <\phi_0,\phi_0> & <\phi_0,\phi_1> & <\phi_0,\phi_2> \\ <\phi_0,\phi_1> & <\phi_1,\phi_1> & <\phi_1,\phi_2> \\ <\phi_0,\phi_2> & <\phi_1,\phi_2> & <\phi_2,\phi_2> \end{bmatrix} = \begin{bmatrix} 4.000 & 6.000 & 9.200 \\ 6.000 & 9.200 & 14.400 \\ 9.200 & 14.400 & 22.996 \end{bmatrix},$$

onde as entradas respectivas foram arredondadas para 3 casas decimais.

2(b) Seja $Q(\phi) = A_0\,\phi(1.6) + A_1\,\phi(1.8)$. Pretende-se determinar a solução do sistema linear $Q(1) = \int_{1.6}^{1.8} dx$ e $Q(x) = \int_{1.6}^{1.8} x\,dx$, isto é,

$$\begin{cases} A_0 + A_1 = 0.2 \\ 1.6\,A_0 + 1.8\,A_1 = 1/2\,(1.8^2 - 1.6^2) = 0.34, \end{cases}$$

de solução $A_0 = A_1 = 0.1$. Assim, trata-se da regra dos trapézios, com passo $h = 0.2$,

$$Q(f) = 0.1\ (f(1.6) + f(1.8)) = h/2\ (f(1.6) + f(1.8)) \ .$$

2(c) A função ϕ satisfaz $\phi \in C^\infty([1.2, 1.6])$. É aplicável a fórmula de erro de interpolação. Assim, existe $\xi \in (1.2, 1.6)$, tal que para $x \in [1, 2, 1.6]$,

$$\phi(x) - p(x) = f^{(3)}(\xi)/6\ (x - 1.2)\,(x - 1.4)\,(x - 1.6) \ .$$

Por conseguinte,

$$\begin{array}{rl} |\phi(1.35) - p(1.35)| & \leq 6/6\,|(1.35 - 1.2)\,(1.35 - 1.4)\,(1.35 - 1.6)| \\ & \leq 0.001875 \\ & < 0.002 \ . \end{array}$$

2(d) Uma vez que p interpola a função em três nós consecutivos, $x_0 = 1.2$, $x_1 = 1.4$ e $x_2 = 1.6$, de espaçamento $h = 0.2$, o valor do integral em causa é dado pela regra de Simpson:

$$\int_{1.2}^{1.6} p(x)\,dx = \frac{h}{3}\left(\phi(1.2) + 4\,\phi(1.4) + \phi(1.6)\right) \simeq -0.440\ .$$

3 (a) Para $f(t,y) = t^2 - \cos(y)$ e $h > 0$, o método é da forma $y_0 = 2$ e

$$y_{i+1} = y_i + \frac{h}{2}\left[f(t_i, y_i) + f(t_i + h, y_i + h\,f(t_i, y_i)\right], \quad i = 0, 1, \ldots\ .$$

Como

$$\begin{aligned} f(t + h, y + h\,f(t,y)) &= (t+h)^2 - \cos\left(y + h\,f(t,y)\right) \\ &= (t+h)^2 - \cos\left(y + h\,(t^2 - cos(y))\right), \end{aligned}$$

tem-se

$$\begin{aligned} &y_0 = 2 \\ &y_{i+1} = y_i + \frac{h}{2}\left[t_i^2 - \cos(y_i) + (t_i + h)^2 - \cos\left(y_i + h\,(t_i^2 - cos(y_i))\right)\right],\ i = 0, 1, \ldots\ . \end{aligned}$$

3(b)

$$\begin{aligned} &t_0 = 1, y_0 = 2, \quad h = 0.2 \\ &y_1 = 2 + 0.1\left(1 - \cos(2) + 1.2^2 - \cos(2 + 0.2(1 - \cos(2))\right), \\ &\text{donde} \\ &y_1 \simeq 2.35098238016\ . \end{aligned}$$

3(c) Em $[1, T] = [1, 2]$, seja $h = (T-1)/N$, para $N \geq 1$. O erro global é da ordem de h^2, isto é, existe constante $c > 0$, tal que

$$E_h = max_{0 \leq i \leq N} |y(t_i) - y_i| \leq c\,h^2\ .$$

Consequentemente $\lim_{h \to 0} E_h = 0$.

A.2.35 (Teste – 11 de Abril 2019)

1(a) Diga o que entende por unidade de arredondamento de um sistema de ponto flutuante.

1(b) Qual é a unidade de arredondamento do sistema $FP(10, 4, -20, 20)$, no caso de arredondamento simétrico?

1(c) Se no sistema $FP(10, 4, -20, 20)$ realizar a operação $\sqrt{3} - 172/99$, qual o valor máximo do erro relativo do resultado que vai obter? Justifique, com base nas fórmulas de propagação do erro.

2) Seja $k \in \mathbb{R}^+$. Considere a sucessão de números reais definida por

$$x_0 = 0.5, \qquad x_{n+1} = g(x_n), \quad n = 0, 1, \ldots, \quad \text{onde} \quad g(x) = k\,x\,(0.5 - x)\ .$$

(a) Sendo $k < 8$, poderá afirmar que todos os termos desta sucessão pertencem ao intervalo $[0, 0.5]$? Justifique.

(b) Quantos pontos fixos tem a função g? Determine-os exactamente.

(c) No caso de $k = 5$ classifique cada ponto fixo (atractor, repulsor ou neutro).

(d) Seja $k = 5$ e $x_0 = 0.25$. Se pretendesse aproximar o ponto fixo atractor, utilizando o método iterativo definido pela função g, quantas iterações teria que realizar para garantir um erro absoluto inferior a 10^{-5} ? (Justifique sem efectuar iterações).

(e) Ainda no caso de $k = 5$, justifique que é possível aproximar ambos os ponto fixos z_1 e z_2 de g usando o método de Newton. Indique números a e b tais que, se $x^{(0)} = a$ o método converge para z_1, e se $x^{(0)} = b$, converge para z_2. Qual é nesse caso a função iteradora?

3) Seja $a \in \mathbb{R}$, $a \neq 0$. Considere a seguinte fórmula iteradora:

$$\begin{aligned} x_1^{(n+1)} &= \frac{2 + x_2^{(n)}}{3} \\ x_2^{(n+1)} &= \frac{1 + x_1^{(n)} + x_3^{(n)}}{3}, \qquad n = 0, 1, \ldots \\ x_3^{(n+1)} &= \frac{-1 + a + x_2^{(n)}}{a} \end{aligned}$$

(a) Sabendo que se trata de um método iterativo para um sistema linear, diga qual o método e indique a matriz do sistema.

(b) Justifique que o método é convergente se e só se $a > 3/8$ ou $a < -3/10$.

(c) Seja $a = 2$. Considerando $x^{(0)} = (1, 1/2, 1)$, efectue a primeira iteração do método considerado e obtenha um majorante de $\|x - x^{(1)}\|_\infty$.

Resolução

1(a) A unidade de arredondamento é o valor máximo que pode atingir o erro de arredondamento relativo nesse sistema.

1(b) $u = 0.5 \times 10^{1-4} = 0.5 \times 10^{-3}$.

1(c) Seja $x = \sqrt{3}$ e $y = 172/99$. Já vimos que a unidade de arredondamento é $u = 0.5 \times 10^{-3}$. Logo, o erro de arredondamento relativo dos dados não excede esse valor, isto é, $|\delta x| \leq u$, $|\delta y| \leq u$. Seja $z = x - y$. O erro relativo de $\bar{z}$ satisfaz a desigualdade

$$|\delta \bar{z}| \leq \frac{|x|\,|\delta x| + |y||\delta y|}{|x - y|} \leq \frac{(|x| + |y|)\,u}{|x - y|} \simeq 0.3259.$$

2 (a) Visto que $g(0.5) = 0$ e $g(0) = 0$ (0 é um ponto fixo de g) todos os termos da sucessão são 0 ou 0.5 .

(b) É necessário resolver a equação $g(z) = z$, ou seja, $k\,z(0.5 - z) = z$. Tratando-se de uma equação quadrática, tem duas raízes. A primeira é $z_1 = 0$. Para a segunda, resulta

$$k(0.5 - z) = 1 \Leftrightarrow 0.5k - 1 = kz \Leftrightarrow z = 0.5 - \frac{1}{k} \ .$$

Assim, os pontos fixos são $z_1 = 0$, $z_2 = 0.5 - 1/k$.

(c) Para $k = 5$, de acordo com a alínea anterior, $z_1 = 0$, $z_2 = 0.5 - 0.2 = 0.3$. Logo $g'(z_1) = 0.5\,k = 2.5$, pelo que z_1 é repulsor. No caso de z_2, $g'(z_2) = 0.5\,k - 2k\,z_2 = 2.5 - 3 = -0.5$. Assim, z_2 é atractor.

(d) Para $k = 5$, já vimos que o ponto fixo atractor é $z_2 = 0.3$. Verifiquemos que no intervalo $I = [0.25, 0.32]$ (por exemplo) estão satisfeitas as condições do teorema do ponto fixo. Temos

$$g(0.25) = 0.3125 \in I; \quad g(0.32) = 0.288 \in I.$$

A derivada de g satisfaz

$$g'(0.25) = 0; \quad g'(0.32) = -0.7 \ .$$

Visto que g' é monótona decrescente ($g''(x) = -10$), concluímos que esta função é não positiva em I, logo g é decrescente em I. Por conseguinte $g(I) \subset I$.
Atendendo a que $\max_{x\in I} |g'(x)| = |g'(0.32)| = 0.7 < 1$, pelo teorema do ponto fixo está garantida a convergência da sucessão para $z_2 = 0.3$, quando $x_0 = 0.25$. Além disso, de acordo com a estimativa do erro, temos

$$|z - x_k| < L^k |z - x_0|, \quad \forall k \in N,$$

onde $L = 0.7$ e e $|z - x_0| = 0.05$. Por conseguinte, para que se verifique $|z - x_k| < 10^{-6}$, basta que $k > \log_L \left(\dfrac{10^{-6}}{0.05}\right) = 23.88$. Assim, efectuando 24 iterações é satisfeito o critério de erro considerado.

(e) Para determinar os pontos fixos de g é necessário resolver a equação $kz(0.5-z) = z$. Para aplicar o método de Newton podemos utilizar a função iteradora

$$g(x) = x - f(x)/f'(x),$$

onde $f(x) = 5\,x\,(0.5 - x) - x$. Note-se que $f'(x) = 1.5 - 10\,x$, pelo que $f'(x) = 0$ sse $x = 0.15$. Logo a função iteradora não está definida neste ponto.
Para aproximar o ponto fixo $z_2 = 0$ devemos utilizar uma aproximação inicial x_0 menor que 0.15. Verifiquemos que, com $x_0 \in [-0.1, 0.1]$, estão satisfeitas as condições de convergência do método de Newton:

$f(-0.1)f(0.1) < 0$,
$f'(x) \neq 0, \quad \forall x \in [-0.1, 0.1]$,
$f''(x) = -10$,
$f(-0.1) = -0.2 < 0$,

de onde resulta que $f(-1)f''(x) > 0$. Logo, para $x_0 = -0.1$, o método converge para $z = 0$.

Para aproximar o ponto fixo $z_2 = 0.3$ devemos utilizar uma aproximação inicial x_0 maior que 0.15. Verifiquemos que, por exemplo, no intervalo $[0.25, 0.32]$, estão satisfeitas as condições de convergência do método de Newton:
$f(0.25)f(0.32) < 0$,
$f'(x) \neq 0, \quad \forall x \in [0.25, 0.32]$,
$f''(x) = -5$,
$f(0.32) = -0.032 < 0$, de onde resulta que $f(0.32)f''(x) > 0$.

Para $x_0 = 0.32$, estão satisfeitas as condições de convergência pelo que o método converge para $z_2 = 0.3$.

3 (a) Dado que no segundo membro da fórmula iteradora só aparecem as componentes da iterada $x^{(n)}$, trata-se do método de Jacobi.
As componentes da matriz A do sistema linear podem ser deduzidas através dos coeficientes da fórmula iteradora. Desta fórmula resulta que o sistema $A\,x = b$ que se pretende resolver é da forma

$$\begin{aligned} x_1 &= \frac{2 + x_2}{3} \\ x_2 &= \frac{1 + x_1 + x_3}{3} \\ x_3 &= \frac{-1 + a + x_2}{a} \end{aligned} \quad ,$$

o que equivale a

$$\begin{cases} 3\,x_1 - x_2 & = 2 \\ -x_1 + 3\,x_2 - x_3 & = 1 \\ -x_2 + a\,x_3 & = -1 + a \ . \end{cases}$$

Logo a matriz do sistema é

$$A = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & a \end{bmatrix} .$$

(b) Comecemos por determinar a matriz de iteração do método de Jacobi. Da fórmula iteradora resulta que

$$C_J = \begin{bmatrix} 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{a} & 0 \end{bmatrix} .$$

Calculando os valores próprios de C_J, tem-se que $\lambda_1 = 0$, $\lambda_{2,3} = \pm\dfrac{\sqrt{3+a}}{3\sqrt{a}}$. Por conseguinte,

$$\rho(C_J) = \left| \frac{\sqrt{3+a}}{3\sqrt{a}} \right| .$$

Assim, a condição necessária e suficiente de convergência do método é que

$$\left| \frac{\sqrt{3+a}}{3\sqrt{a}} \right| < 1,$$

donde $a < -3/10$ ou $a > 3/8$.

(c) Temos

$$\begin{aligned} x_1^{(1)} &= \frac{2 + x_2^{(0)}}{3} = \frac{5}{6} \\ x_2^{(1)} &= \frac{1 + x_1^{(0)} + x_3^{(0)}}{3} = 1 \\ x_3^{(1)} &= \frac{-1 + a + x_2^{(0)}}{a} = \frac{3}{4} \ . \end{aligned}$$

Logo, $\|x^{(0)} - x^{(1)}\|_\infty = 1/2$. Apliquemos a fórmula do erro do método de Jacobi,

$$\|x - x^{(1)}\|_\infty \leq \frac{\|C_J\|_\infty}{1 - \|C_J\|_\infty} \|x^{(0)} - x^{(1)}\|_\infty \ .$$

Visto que, de acordo com a alínea anterior, $\|C_J\|_\infty = 2/3$, tem-se

$$\|x - x^{(1)}\|_\infty \leq 1 \ .$$

A.2.36 (Teste – 29 Maio 2019)

1) Dado o sistema *não linear*, nas incógnitas a, b, c, d, [1.5]

$$4\,a + b^2 + 2 = 0, \quad 2\,b^4 + c = d, \quad a + 2\,c^3 = b^2, \quad a^2 - 2\,d^3 = \sin(c),$$

a partir de $x^{(0)} = (1, 1, \pi, 1)$, indique o *sistema linear* que permite obter a primeira iteração do método de Newton. (Não se pede para calcular a solução deste sistema).

2) Sejam $f(x) = |x + 2|$, e os pontos $x_i = 1 + i/5$, para $i = 0, 1, \ldots, 4$.

(a) Determine p, polinómio interpolador nos nós x_0, x_2 e x_4 . [1.0]

(b) Obtenha $|f(x_3) - p(x_3)|$, e diga se o resultado calculado está de acordo com a teoria. Justifique. [1.0]

(c) Determine α e $\beta \in \mathbb{R}$ que minimizam $F(\alpha, \beta) = \sum_{k=0}^{2} (f(x_k) - \alpha - \beta\, x_k)^2$. (Pretende-se que indique detalhadamente como obtém um sistema linear a partir do qual se calcula α e β. Não é preciso determinar a solução do sistema). [1.0]

3) Para aproximar o integral $\displaystyle\int_{-1}^{1} f(x)\,dx$, considere a regra dos trapézios $T(f)$, com dois subintervalos, bem como a regra

$$Q(f) = f(-1) + f(1) + \frac{1}{3}\left(f'(-1) - f'(1)\right) \ .$$

(a) Determine o grau da regra Q(f). Justifique. [1.5]

(b) Sendo $f(x) = 2\,x^4$, obtenha o erro exacto das regras $T(f)$ e $Q(f)$. [1.0]

(c) Para $f(x) = \sin(x - 1)$, quantas subdivisões do intervalo $[-1, 1]$ deverá considerar para que o erro da regra de Simpson seja inferior a 10^{-3}? Justifique. [1.5]

[1.5] **4)** Considere o problema de valor inicial $y'(t) = 5\,t^4$, $y(0) = -1$, para $0 \leq t \leq 1$. Escreva a equação às diferenças do método de Taylor de segunda ordem, com passo $h = 0.2$. Calcule aproximação de $y(0.4)$ mediante aplicação do método.

Resolução

1) Para $f(a,b,c,d) = (4\,a + b^2 + 2, 2\,b^4 + c - d, a + 2\,c^3 - b^2, a^2 - 2\,d^3 - \sin(c))^T$, tem-se

$$J_f(a,b,c,d) = \begin{bmatrix} 4 & 2\,b & 0 & 0 \\ 0 & 8\,b^3 & 1 & -1 \\ 1 & -2\,b & 6\,c^2 & 0 \\ 2\,a & 0 & -\cos(c) & -6\,d^2 \end{bmatrix} .$$

Como $x^{(0)} = (1,1,\pi,1)^T$, resulta $f(x^{(0)}) = (7, 1+\pi, 2\,\pi^3, -1)^T$. O sistema a resolver é da forma $J_f(x^{(0)})\,\Delta x^{(0)} = -f(x^0)$, isto é,

$$\begin{bmatrix} 4 & 2 & 0 & 0 \\ 0 & 8 & 1 & -1 \\ 1 & -2 & 6\,\pi^2 & 0 \\ 2 & 0 & 1 & -6 \end{bmatrix} \begin{bmatrix} \Delta^{(0)}\,a \\ \Delta^{(0)}\,b \\ \Delta^{(0)}\,c \\ \Delta^{(0)}\,d \end{bmatrix} = \begin{bmatrix} -7 \\ -1-\pi \\ -2\,\pi^3 \\ 1 \end{bmatrix} .$$

2(a)

x_i	f_i	$f[..]$	$f[...]$
$x_0 = 1$	3		
		1	
$x_2 = 7/5$	$17/5$		0
		0	
$x_4 = 9/5$	$19/5$		

$$p(x) = 3 + (x-1) = 2 + x = f(x) .$$

2(b) $|f(x_3) - p(x_3)| = 0$. Com efeito, como $f \in C^2([x_0, x_4])$ e $f''(x) = 0$, da fórmula teórica do erro de interpolação resulta que o erro de interpolação é nulo para qualquer valor x no intervalo, em particular para x_3. De facto,

$$f(x_3) = |x_3 + 2| = 3.6,\ p(x_3) = x_3 + 2 = 3.6 \Longrightarrow f(x_3) - p(x_3) = 0 .$$

2(c) Atendendo a que

$$\begin{array}{ll} x_0 = 1 & f(1) = 3 \\ x_1 = 6/5 & f(6/5) = 16/5 \\ x_2 = 7/5 & f(7/5) = 17/5, \end{array}$$

como $F(\alpha,\beta) = \sum_{k=0}^{2} (f(x_k) - (\alpha + \beta x_k))^2$, o mínimo pretendido é atingido se $\partial F(\alpha,\beta)/\partial\alpha = 0$ e $\partial F(\alpha,\beta)/\partial\beta = 0$, isto é,

$$\begin{cases} \sum_{k=0}^{2} \alpha + \beta x_k - f(x_k) & = 0 \\ \\ \sum_{k=0}^{2} (\alpha + \beta x_k - f(x_k))\,x_k & = 0, \end{cases}$$

ou seja,

$$\begin{cases} 3\,\alpha & + \left(\sum_{k=0}^{2} x_k\right)\,\beta & = \sum_{k=0}^{2} f(x_k) \\ \left(\sum_{k=0}^{2} x_k\right)\,\alpha & + \left(\sum_{k=0}^{2} x_k^2\right)\,\beta & = \sum_{k=0}^{2} f(x_k)\,x_k\ . \end{cases}$$

3(a) A regra é de grau k se for exacta para os monómios 1, x, ..., x^k, mas não é exacta para x^{k+1}. Ora

$$\begin{array}{ll} Q(1) = 1 & \text{e}\quad I(1) = \int_{-1}^{1} dx = 2 \\ Q(x) = 1/3\,(1-1) = 0 & \text{e}\quad I(1) = \int_{-1}^{x} dx = 0 \\ Q(x^2) = 2 + 1/3(-2-2) = 2/3 & \text{e}\quad I(1) = \int_{-1}^{1} x^2\,dx = 2/3 \\ Q(x^3) = 1/3\,(3-3) = 0 & \text{e}\quad I(1) = \int_{-1}^{1} x^3\,dx = 0 \\ Q(x^4) = 2 + 1/3(-4-4) = -2/3 & \text{e}\quad I(1) = \int_{-1}^{1} x^4\,dx = 2/5\ . \end{array}$$

Por conseguinte, a regra é de grau 3.

3(b) Para $f(x) = 2\,x^4$, tem-se $Q(f) = 2 + 2 + 1/3\,(-8-8) = -4/3$ e $I(f) = 4/5$. Assim,

$$I(f) - Q(f) = 4/5 + 4/3 = 32/15\ .$$

Quanto à regra dos trapézios com $h = 1$, obtém-se

$$T(f) = h/2\ (f(-1) + 2f(0) + f(1)) = 1/2\,(2+2) = 2\ .$$

Logo, $I(f) - T(f) = 4/5 - 2 = -6/5$.

3(c) Seja $\epsilon = 10^{-3}$. Para $f(x) = \sin(x-1)$, tem-se $f^{(4)}(x) = \sin(x-1) = f(x)$ e $f^{(5)}(x) = \cos(x-1)$. Atendendo à fórmula de erro para a regra de Simpson, com $N \geq 2$ subintervalos, resulta

$$|I(f) - S_N(f)| \leq \frac{b-a}{180}\,\frac{(b-a)^4}{N^4}\,max_{a\leq x\leq b}|f^{(4)}(x)|$$

Ora, como $f^{(5)}(x) = 0 \Longrightarrow x - 1 = \pi/2$, isto é, para $x \in [-1, 1]$, $f^{(4)}$ tem extremo em $x = \pi/2 + 1$, donde $max_{a\leq x\leq b}|f^{(4)}(x)| = |\sin(1 + \pi/2 - 1)| = 1$. Assim,

$$\frac{2^5}{180}\,\frac{1}{N^4} < \epsilon \iff N > \left(\frac{2^5}{180\,\epsilon}\right)^{1/4} \simeq 3.65\ .$$

Logo, $N = 4$.

4) Como $f(t,y) = 5\,t^4$, $\partial f/\partial t(t,y) = 20\,t^3$, e $\partial f/\partial y(t,y) = 0$, a equação às diferenças do método escreve-se

$$\begin{array}{ll} y_0 & = -1, \\ y_{i+1} & = y_i + 5\,h\,t_i^4 + \dfrac{h^2}{2}\,\left(20\,t_i^3\right), \\ & \text{isto é,} \\ y_{i+1} & = y_i + 5\,h\,t_i^3\,(t_i + 2\,h), \quad \text{com } t_i = i\,h, \text{ para } i = 0, 1, \ldots\ . \end{array}$$

Primeiro passo: $t_0 = 0,\ y_0 = -1 \quad y_1 = y_0 = -1;$

Segundo passo: $t_1 = h = 0.2,$
$y_2 = -1 + 0.2^3\, t_1^3\ (0.2 + 0.4) = -0.9952 \simeq y(0.2)\ .$

A.2.37 (Teste – 18 Dezembro 2019)

1) Em $[0,2] \times [0,2]$, o sistema não linear $\begin{cases} 4\,x^2 - 3/2\,y^2 = 1 \\ 3\,x^2 + y^2 - 1 = 0 \end{cases}$ tem uma solução $Z = (\sqrt{5/17}, \sqrt{2/17})^T$.

[1.5] Efectuando cálculos exactos determine $||Z - W^{(1)}||_1$, sendo $W^{(1)}$ a primeira iterada do método de Newton, iniciado com $W^{(0)} = (1,1)^T$.

2) Dada a função $f(x) = \ln(x+2)$, considere a tabela de valores $(x_i, f(x_i))$, onde $x_i = -1 + i$, com $i = 0, 1, 2, 3$.

[1.0] **(a)** Sabendo que $f[-1,0,1,2] = \ln(32/27)/6$, obtenha uma expressão exacta para o polinómio interpolador dos 4 pontos tabelados. Justifique.

[1.5] **(b)** Sendo $q(x)$ o polinómio interpolador nos 3 primeiros nós, mostre que existe um valor $\theta \in (-1,1)$, tal que $\ln(5/2) = q(1/2) - 1/[8\,(\theta+2)^3]$. Justifique esta última igualdade com base num resultado teórico que conheça.

[1.5] **(c)** Suponha que a função f é aproximada, no sentido dos mínimos quadrados, por funções do tipo $\Psi(x) = a_0 + a_1(x+2) + a_2(x+2)^2$, nos pontos x_i considerados. Determine a matriz do sistema de equações normais para o cálculo de a_0, a_1 e a_2. Apresente justificação para que tal sistema seja referido como de "equações normais".

3) Para aproximar $I(g) = \displaystyle\int_{-1}^{1} g(x)\,dx$, considere a regra de quadratura
$Q(g) = a_1\, g(x_1) + a_2\, g(1), \quad$ onde $\quad x_1 \neq 1, \quad a_1, a_2 \neq 0$

[1.5] **(a)** Aplicando o método dos coeficientes indeterminados, determine x_1 e os pesos a_1, a_2, de modo que a regra seja exacta para polinómios de grau ≤ 2.

[1.0] **(b)** Qual o grau de precisão da regra que obteve na alínea anterior? Justifique.
(Caso não tenha resolvido a alínea anterior diga, justificando, qual é o grau da regra dos trapézios.)

[1.0] **(c)** Sendo $g(x) = 2\,x^3 + x + 1$, é verdade que a regra de Simpson composta, com 11 nós, é exacta? Justifique.

[1.0] **4)** Considere o problema de valor inicial $y' = e^y + x$, $y(1) = 0$, para $1 \leq x \leq 2$. Obtenha uma aproximação de $y(1.2)$ aplicando o método de Euler explícito, com passo $h = 0.1$. Comece por escrever a equação às diferenças do método.

Resolução

1) A primeira iterada do método pode obter-se resolvendo o sistema linear

$$J_f(W^{(0)})\,\Delta W^{(0)} = -f(W^{(0)}),$$

donde

$$W^{(1)} = W^{(0)} + \Delta W^{(0)} \ .$$

Assim,

$$\left[\begin{array}{cc} 8 & -3 \\ 6 & 2 \end{array}\right] \left[\begin{array}{c} \Delta w_1 \\ \Delta w_2 \end{array}\right]^{(0)} = -\left[\begin{array}{c} 3/2 \\ 3 \end{array}\right] = \left[\begin{array}{c} -3/2 \\ -3 \end{array}\right]$$

$$\Delta w_1^{(0)} = \frac{\left|\begin{array}{cc} -3/2 & -3 \\ -3 & 2 \end{array}\right|}{34} = \frac{-6}{17}$$

$$\Delta w_2^{(0)} = \frac{\left|\begin{array}{cc} 8 & -3/2 \\ 6 & -3 \end{array}\right|}{34} = \frac{-15}{34}$$

$$W^{(1)} = \left[\begin{array}{c} 1 \\ 1 \end{array}\right] + \left[\begin{array}{c} -6/17 \\ -15/34 \end{array}\right] = \left[\begin{array}{c} 11/17 \\ 19/34 \end{array}\right] \ .$$

Donde,

$$||Z - W^{(1)}||_1 = |\sqrt{5/7} - 11/17| + |\sqrt{2/17} - 19/34| \simeq 0.414 \ .$$

2(a) Da tabela de diferenças divididas,

x_i	$f(x_i)$	$f[..]$	$f[...]$
-1	0		
		$\ln(2)$	
0	$\ln(2)$		$\dfrac{\ln(3) - 2\,\ln(2)}{2}$
		$\ln(3) - \ln(2)$	
1	$\ln(3)$		

obtém-se o polinómio interpolador nos 3 primeiros nós,

$$q(x) = \ln(2)(x+1) + \frac{\ln(3) - 2\ln(2)}{2}\,(x+1)\,x \ .$$

O polinómio interpolador dos 4 nós é

$$p(x) = q(x) + \frac{\ln(32/27)}{6}\,(x+1)\,x\,(x-1) \ .$$

2(b) Como $f \in C^3([-1,1])$ e $f^{(3)}(x) = 2/(x+2)^3$, para $x = 1/2$, sabe-se que existe $\theta \in (-1,1)$, tal que o erro de interpolação em x satisfaz a igualdade $f(x) - q(x) = f^{(3)}(\theta)/6\,(x+1)\,x\,(x-1)$, isto é,

$$\ln(5/2) = q(1/2) + 2/(6\,(\theta+2)^3) \times 3/2 \times 1/2 \times (-1/2),$$

igualdade que coincide com a dada.

2(c) Seja $\phi(x) = a_0\,\phi_0(x) + a_1\,\phi_1(x) + a_2\,\phi_2(x)$, com $\phi_0(x) = 1, \phi_1(x) = x - 2$ e $\phi_2(x) = (x-2)^2$.
De

x_i	-1	0	1	2	
$\phi_0(x_i) = 1$	1	1	1	1	$\leftarrow \phi_0$
$\phi_1(x_i) = x_i - 2$	1	2	3	4	$\leftarrow \phi_1$
$\phi_2(x_i) = (x_i-2)^2$	1	4	9	16	$\leftarrow \phi_2$

resulta

$$
\begin{aligned}
&< \phi_0, \phi_0 >= 4\\
&< \phi_0, \phi_1 >= 1+2+3+4 = 10\\
&< \phi_0, \phi_2 >= 1+4+9+16 = 30\\
&< \phi_1, \phi_1 >= 1+4+9+16 = 30\\
&< \phi_1, \phi_2 >= 1+8+27+64 = 100\\
&< \phi_2, \phi_2 >= 1+16+81+16^2 = 354
\end{aligned}
$$

A matriz do sistema de equações normais é da forma

$$\begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix} .$$

Sendo $f-g = f-(a_0\phi_0 + a_1\,phi_1 + a_2\,\phi_2)$, os valores a_0, a_1, a_2 que minimizam $||f-g||_2^2$ satisfazem as condições de ortogonalidade $\langle f-g, \phi_0\rangle = 0$, $\langle f-g, \phi_1\rangle = 0$ e $\langle f-g, \phi_2\rangle = 0$, as quais se traduzem no sistema linear de incógnitas a_0, a_1, a_2 cuja matriz é a que foi construída acima.

3(a) A regra é exacta para qualquer polinómio de grau ≤ 2 se e só se é exacta para os monómios $1, x, x^2$. Ou seja,

$$\begin{cases} a_1 + a_2 & = 2 \\ x_1\,a_1 + a_2 & = 0 \\ x_1^2\,a_1 + a_2 & = 2/3 \end{cases}$$

Das duas primeiras equações resulta,

$$(x_1 - 1)\,a_1 = -2 \qquad \Longleftrightarrow \qquad a_1 = \frac{-2}{x_1 - 1} \quad (*)$$

Das duas últimas equações obtém-se,

$$x_1^2\,a_1 - x_1\,a_1 = 2/3, \quad \text{logo} \quad x_1\,a_1\,(x_1-1) = 2/3 \quad (**)$$

Como $x_1 - 1 = -2/a_1$, de $(**)$ vem,

$$-2\,x_1 = 2/3 \quad \Longleftrightarrow \quad x_1 = \frac{-1}{3}, \quad \text{donde por (*)} \quad a_1 = -2/(x_1-1) = 3/2\ .$$

Da primeira equação sabemos que $a_2 = 2 - a_1$. Por conseguinte, $a_2 = 2 - 3/2 = 1/2$. Assim,

$$Q(g) = 3/2\,g(-1/3) + 1/2\,g(1)$$

é regra de grau ≥ 2.

3(b) Como $Q(x^3) = 3/2\,(-1/27) + 1/2 = 4/9$ e $I(x^3) = 2/4 = 1/2 \neq Q(x^3)$, a regra em causa é de grau 2.

3(c) Uma vez que a regra de Simpson (simples ou composta) é de grau 3, ela é exacta para qualquer polinómio de grau ≤ 3 e, em particular, para o polinómio dado.

4) Para $f(x,y) = e^y + x$ e $h = 0.1$, a equação às diferenças do método de Euler escreve-se:

$$\begin{array}{ll} y_0 & = 0 \\ y_{i+1} & = y_i + 0.1\,(e^{y_i} + x_i)\,, \quad i = 0, 1, \ldots \\ \text{isto é,} & \\ y_{i+1} & = y_i + 0.1\,(e^{y_i} + 1 + 0.1\,i)\,, \qquad i = 0, 1, \ldots \end{array}$$

Por conseguinte, para $x_0 = 1$,

$$\begin{array}{l} y_1 = 0.1\,(e^0 + 1) = 0.2 \simeq y(1.1) \\ y_2 = 0.2 + 0.1\,(e^{0.2} + 1 + 0.1) = 0.432140275\cdots \simeq y(1.2)\ . \end{array}$$

A.2.38 (Exame, Parte 1 – 21 Janeiro 2020)

1) Considere a função $f(x) = \sin(x-1)/x$, para $0 < x \leq 1$.

(a) Diga, justificando, se a função é bem condicionada para valores próximos de $x = 1$. [1.0]

(b) Sendo $\bar{x} = 0.999$ um número aproximado representado num sistema decimal com 3 dígitos na mantissa, calcule uma estimativa do quociente de erros relativos $|\delta_{f(\bar{x})}|/|\delta_{\bar{x}}|$, e conclua se o valor que obteve está ou não de acordo com a alínea anterior. [1.0]

2) Sabe-se que a equação $x/3 = \cos(x)$ tem uma única raiz $z \in I = [1, 3/2]$. Para uma função contínua $\phi(x) \neq 0,\ \forall x \in I$, considere a função iteradora

$$g(x) = x + \phi(x)\ (x - 3\,\cos(x))\ .$$

(a) Admitindo que o processo iterativo $x_{n+1} = g(x_n),\ n = 0, 1, \ldots$ é convergente para um número $\beta \in I$, mostre que β é raiz da equação dada. [1.0]

(b) Para $\phi(x) = -1/4$, prove que o método de ponto fixo converge para a raiz z, qualquer que seja $x_0 \in I$. A convergência é linear? Justifique. [1.5]

(c) Fazendo $x_0 = 1$, obtenha x_3 mediante aplicação do método de Newton. Calcule uma estimativa do erro $z - x_3$. [1.5]

(d) Para o método da alínea anterior sabe-se que $\lim_{n\to\infty} \left(|z - x_{n+1}|/|z - x_n|^2\right) = c > 0$. Obtenha uma aproximação do valor da constante c. Justifique. [1.0]

3) Considere o sistema linear de quatro equações: $x_1+x_2/2-x_3/3=0$, $-x_1+2\,x_2=0$, $-x_2+3\,x_3=0$ e $-x_1+5\,x_4=0$.

(a) Partindo de $x^{(0)}\neq(0,0,0,0)^T$, diga se o método de Jacobi converge para a solução do sistema. Justifique. [1.0]

(b) Ainda para o método da alínea anterior, sendo $x^{(0)}=(1,1,1,1)^T$, mostre que $||x^{(k)}||_\infty\leq(5/6)^k$, para $k=1,2,\ldots$. [1.0]

[1.0] **(c)** Escreva as fórmulas iterativas do método de Gauss-Seidel e calcule $||x^{(1)}-x^{(0)}||_1$, sendo $x^{(0)}=(1,0,0,-1)^T$.

Resolução

1(a) Sabemos que $\delta_{f(\bar{x})}\simeq P_f(x)\,\delta_{\bar{x}}$. Como

$$f'(x)=\frac{\cos(x-1)\,x-\sin(x-1)}{x^2},$$

tem-se, para $f(x)\neq 0$, e $x>0$,

$$\begin{aligned}P_f(x) &= \frac{x\,f'(x)}{f(x)}=\frac{\cos(x-1)\,x-\sin(x-1)}{\sin(x-1)},\quad x>0\\ &= \frac{x\,\cos(x-1)}{\sin(x-1)}-1\ . \qquad\qquad (*)\end{aligned}$$

Por conseguinte,

$$\lim_{x\to 1}|P_f(x)|=+\infty,$$

pelo que a função é mal condicionada para valores de x no intervalo considerado, próximos do extremo superior.

1(b) Atendendo à expressão $(*)$, resulta

$$\frac{|\delta_{f(\bar{x})}|}{|\delta_{\bar{x}}|}\simeq\frac{|\bar{x}\,\cos(\bar{x}-1)|}{|\sin(\bar{x}-1)|}\simeq 999,$$

o que está de acordo com o mau condicionamento previsto na alínea anterior.

2(a) Dado que

$$\lim_{n\to\infty}x_{n+1}=\beta=g(\lim_{n\to\infty}x_n)=g(\beta)$$

(onde a última igualdade é válida porquanto g é contínua), tem-se que

$$\beta=\beta+\phi(\beta)\,(\beta-3\,\cos(\beta))\ .$$

Como $\phi(\beta)\neq 0$, obtém-se

$$\beta-3\,\cos(\beta)=0,$$

ou seja, β é raiz da equação dada.

2(b) Verifiquemos que são válidas as condições do teorema do ponto fixo, para a função iteradora

$$g(x) = x - 1/4\ (x - 3\ \cos(x))\,, \quad I = [1, 3/2], \quad g \in C^1(I)\ .$$

Tem-se,

$$\begin{array}{l} g'(x) = 1 - 1/4\,(1 + 3\ \sin(x)) \\ g''(x) = -3/4\ (\cos(x)) < 0 \quad \forall x \in I\ . \end{array}$$

Assim, a função g' é positiva e decrescente no intervalo. Como

$$g'(1) \simeq 0.119, \quad g'(3/2) \simeq 0.0019,$$

resulta que $L = max_{x\in I}|g'(x)| = g'(1) \simeq 0.0019$. Além disso, a função g é positiva e crescente. Por conseguinte, atendendo a que $g(1) \simeq 1.115$ e $g(3/2) \simeq 1.178$, tem-se

$$1 < g(1) \leq g(x) \leq g(3/2) < 3/2, \quad \forall x \in I,$$

ou seja, $g(I) \subset I$. Pelo referido teorema sabemos que existe um único ponto fixo z de g em I; o método de ponto fixo converge para o ponto fixo z, qualquer que seja o valor inicial $x_0 \in I$ considerado; e o ponto fixo é a raiz z (única) da equação dada, no intervalo em causa.
A convergência é linear pois $0 < |g'(z)| < 1$, e sabemos que

$$\lim_{k\to\infty} (|z - x_{k+1}|/|z - x_k|)) = |g'(z)|\ .$$

2(c) Para $f(x) = x/3 - \cos(x)$, $x \in I$, tem-se $f'(x) = 1/3 + \sin(x)$. A função iteradora de Newton é da forma $g(x) = x - f(x)/f'(x)$. Assim para $x_0 = 1$,

$$\begin{array}{l} x_1 = x_0 - f(x_0)/f'(x_0) \simeq 1.17617314 \\ x_2 = x_1 - f(x_1)/f'(x_1) \simeq 1.17012658 \\ x_3 = x_2 - f(x_2)/f'(x_2) \simeq 1.17012095 \\ x_4 = x_3 - f(x_3)/f'(x_3) \simeq 1.17012095\ . \end{array}$$

Como a convergência é supralinear, podemos estimar o erro de x_3 usando os valores de x_4 e x_3:

$$e_3 = z - x_3 \simeq x_4 - x_3,$$

donde $|e_3| < 10^{-8}$.

2(d) Como $z \simeq x_3$, sabe-se que

$$c = \frac{|f''(z)|}{2\,|f'(z)|} \simeq \frac{|f''(x_3)|}{2\,|f'(x_3)|} \simeq 0.156\ .$$

3(a) A matriz do sistema dado tem diagonal estritamente dominante por linhas. Por conseguinte, o método converge para a solução $(0, 0, 0, 0)$, qualquer que seja a aproximação inicial considerada. Com efeito, o método é da forma

$$\left\{\begin{array}{ll} x_1^{(k+1)} & = -1/2\,x_2^{(k)} + 1/3\,x_3^{(k)} \\ x_2^{(k+1)} & = 1/2\,x_1^{(k)} \\ x_3^{(k+1)} & = 1/3\,x_2^{(k)} \\ x_4^{(k+1)} & = 1/5\,x_1^{(k)}, \qquad\qquad k = 0, 1, \ldots \end{array}\right.$$

Por conseguinte, a respectiva matriz de iteração é

$$C = \begin{bmatrix} 0 & -1/2 & 1/3 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 1/5 & 0 & 0 & 0 \end{bmatrix} .$$

Consequentemente,

$$||C||_\infty = max(5/6, 1/2, 1/3, 1/5) = 5/6 < 1,$$

pelo que o método converge para a solução do sistema $x = (0, 0, 0, 0)$, qualquer que seja $x^{(0)} \neq (0, 0, 0, 0)$ considerado.

3(b) Da alínea anterior, tem-se

$$\begin{array}{l} ||C||_\infty = max(5/6, 1/2, 1/3, 1/5) = 5/6, \\ x - x^{(0)} = (-1, -1, -1, -1), \\ ||x - x^{(0)}||_\infty = 1, \\ ||x - x^{(k)}||_\infty \leq ||C||_\infty^k \, ||x - x^{(0)}||_\infty, \quad \text{logo} \end{array}$$

$$||x^k||_\infty \leq (5/6)^k, \quad k = 1, 2, \ldots \ .$$

3(c) Seja $x^{(0)} = (1, 0, 0, -1)^T$. Do sistema dado obtém-se, equivalentemente,

$$\begin{cases} x_1 = -x_2/2 + x_3/3 \\ x_2 = x_1/2 = -x_2/4 + x_3/6 \\ x_3 = x_2/3 = -x_2/12 + x_3/18 \\ x_4 = x_1/5 = -x_2/10 + x_3/15, \end{cases}$$

donde as fórmulas de iteração do método de Gauss-Seidel:

$$\begin{cases} x_1^{(k+1)} = -x_2^{(k)}/2 + x_3^{(k)}/3 \\ x_2^{(k+1)} = x_1^{(k+1)}/2 = -x_2^{(k)}/4 + x_3^{(k)}/6 \\ x_3^{(k+1)} = x_2^{(k+1)}/3 = -x_2^{(k)}/12 + x_3^{(k)}/18 \\ x_4^{(k+1)} = x_1^{(k+1)}/5 = -x_2^{(k)}/10 + x_3^{(k)}/15, \qquad k = 0, 1, 2, \ldots \ . \end{cases}$$

Por conseguinte,

$$\begin{array}{l} x^{(1)} = (0, 0, 0, 0) \\ ||x^{(1)} - x^{(0)}||_1 = ||x^{(0)}||_1 = 1 + 1 = 2 \ . \end{array}$$

A.2.39 (Exame, Parte 2 – 21 Janeiro 2020)

1) Considere a função

$$g(t) = \frac{t-1}{t+2}, \quad t > -2 \ .$$

(a) Determine o polinómio p, interpolador de g nos nós $t_0 = -1$, $t_1 = 1$, $t_2 = 3$ e $t_3 = 4$. [1.0]

(b) Calcule $\gamma = |g(0) - p(0)|$. Justifique se a fórmula de erro é aplicável e, no caso positivo, compare γ com a estimativa de erro. Comente. [1.5]

2(a) Determine os valores α e $\beta \in \mathbb{R}$ que minimizam [1.5]

$$F(\alpha, \beta) = \sum_{j=-1}^{j=1} \left(\alpha + \beta\, j^2 - \frac{1}{j+2} \right)^2 \ .$$

Escreva todos os cálculos que efectuar.

2(b) Poderá dizer que $F(x, y) \geq 2/9, \quad \forall x, y \in \mathbb{R}$? Justifique. [1.0]

3) Considere o integral $I(f) = \displaystyle\int_0^1 f(x)\, dx$ e a regra $Q(f) = w_1\, f(0) + w_2\, f(1)$, onde $w_1, w_2 \in \mathbb{R}$.

(a) Aplicando o método dos coeficientes indeterminados obtenha os pesos da regra, de modo que $Q(f)$ tenha grau de precisão pelo menos um. Justifique. Qual a designação habitual da regra que determinou? [1.5]

(b) Se aproximar $\displaystyle\int_0^1 \sin(x/3)\, dx$ usando a regra dos trapézios composta, quantos nós de quadratura deverá considerar para que o erro absoluto da aproximação não exceda 10^{-4}? Justifique. [1.5]

4) Considere o problema de valor inicial $y'(t) = 3\, t, \ y(0) = -1$, com $0 \leq t \leq 1$.

(a) Para $N \geq 1$ e $h = 1/N$, escreva a equação às diferenças do método de Euler aplicado ao problema em causa. [0.5]

(b) Mostre que $\lim_{N \to \infty} |y(1) - y_N| = 0$. [1.5]

Resolução

1(a) O polinómio interpolador nas diferenças divididas é da forma

$$p(t) = g(-1) + g[-1, 1]\,(t+1) + g[-1, 1, 3]\,(t+1)(t-1) + g[-1, 1, 3, 4]\,(t+1)(t-1)(t-3) \ .$$

Sabe-se que $g(-1) = -2, g(1) = 0$, $g(3) = 2/5$ e $g(4) = 1/2$. Tem-se,

$$\begin{array}{ll} g[-1,1] & = 2/2 = 1 \\ g[1,3] & = 2/5 \times 1/2 = 1/5 \\ g[3,4] & = 1/2 - 2/5 = 1/10 \end{array}$$

$$g[-1,1,3] = \frac{g[1,3] - g[-1,1]}{4} = -1/5$$

$$g[1,3,4] = \frac{g[3,4] - g[1,3]}{4} = -1/30$$

$$g[-1,1,3,4] = \frac{g[1,3,4] - g[-1,1,3]}{5} = 1/30 \ .$$

Assim,

$$\begin{aligned} p(t) &= -2 + t + 1 - \frac{1}{5}(t+1)(t-1) + \frac{1}{30}(t+1)(t-1)(t-3) \\ &= t - 1 - \frac{1}{5}(t+1)(t-1) + \frac{1}{30}(t+1)(t-1)(t-3) \ . \end{aligned}$$

1(b) Dado que $g \in C^4([-1,4])$, a fórmula teórica do erro de interpolação em $t = 0$ é aplicável. Como

$$g'(t) = \frac{3}{(t+2)^2}, \qquad g^{(2)}(t) = -\frac{6}{(t+2)^3}$$

$$g^{(3)}(t) = \frac{18}{(t+2)^4}, \qquad g^{(4)}(t) = -\frac{72}{(t+2)^5}$$

$$g^{(5)}(t) = -\frac{5 \times 72}{(t+2)^6} > 0 \quad \forall t \in [-1,4],$$

conclui-se que $g^{(4)}$ é sempre negativa e crescente no intervalo em causa. Por conseguinte,

$$M = max_{-1 \le t \le 4} |g^{(4)}(t)| = |g^{(4)}(-1)| = 72 \ .$$

Recorrendo à fórmula de erro, a estimativa de erro de interpolação em $t = 0$ é:

$$\begin{aligned} |g(0) - p(0)| &\le \frac{72}{4!} \left|(0+1)(0-1)(0-3)(0-4)\right| \\ &\le 36 \ . \end{aligned}$$

Uma vez que $g(0) = -1/2$ e $p(0) = -7/10$, resulta $\gamma = |-1/2 + 7/10| = 1/5$. Conclui-se que a estimativa de erro anterior é manifestamente irrealista. Tal acontece devido ao valor elevado de M no intervalo em causa. Além disso, os nós 3 e 4 estão relativamente afastados do ponto $t = 0$ a interpolar, dando como resultado uma ampliação do factor $M/4!$ na fórmula de erro.

2(a) Os valores α e β deverão ser solução do sistema linear $\dfrac{\partial F(\alpha,\beta)}{\partial \alpha} = 0$, $\dfrac{\partial F(\alpha,\beta)}{\partial \beta} = 0$. Isto é,

$$\begin{cases} \displaystyle\sum_{j=-1}^{j=1} \left(\alpha + \beta\, j^2 - \frac{1}{j+2}\right) = 0 \\ \sum_{j=-1}^{j=1} \left(\alpha + \beta\, j^2 - \dfrac{1}{j+2}\right) j^2 = 0, \end{cases}$$

ou seja,

$$\begin{cases} 3\,\alpha + \displaystyle\sum_{j=-1}^{j=1} j^2\,\beta & = \displaystyle\sum_{j=-1}^{j=1} \frac{1}{j+2} \\ \displaystyle\sum_{j=-1}^{j=1} j^2\,\alpha + \sum_{j=-1}^{j=1} j^4\,\beta & = \displaystyle\sum_{j=-1}^{j=1} \frac{j^2}{j+2}\ . \end{cases}$$

Substituindo, tem-se

$$\begin{cases} 3\,\alpha + 2\,\beta & = 11/6 \\ \\ 2\,\alpha + 2\,\beta & = 4/3\ . \end{cases}$$

Cuja solução se obtém facilmente, $\alpha = 1/2$ e $\beta = 1/6$.

2(b) O mínimo da função $F(\alpha,\beta)$ é atingido para $\alpha = 1/2$ e $\beta = 1/6$. Como

$$\begin{aligned} f(1/2, 1/6) &= \sum_{j=-1}^{j=1} \left(\frac{1}{2} + \frac{j^2}{6} - \frac{1}{j+2}\right)^2 \\ &= (1/2 + 1/6 - 1)^2 + (1/2 - 1/2)^2 + (1/2 + 1/6 - 1/3)^2 \\ &= (-1/3)^2 + 0 + (1/3)^2 = \frac{2}{9}, \end{aligned}$$

conclui-se que $F(x,y) \geq 2/9, \quad \forall x, y \in \mathbb{R}$.

3(a) A regra tem grau de precisão ≥ 1 se e só se $Q(1) = I(1)$ e $Q(x) = I(x)$. Isto é,

$$\begin{cases} w_1 + w_2 & = \displaystyle\int_0^1 dx = 1 \\ \\ w_2 & = \displaystyle\int_0^1 x\, dx = 1/2\ . \end{cases}$$

Assim, $w_2 = w_1 = 1/2$, e

$$Q(f) = \frac{1}{2}\,\left[f(0) + f(1)\right]\ .$$

Trata-se da regra dos trapézios simples.

3(b) Para $f(x) = \sin(x/3)$, $f'(x) = 1/3\,\cos(x/3)$ e $f^{(2)}(x) = -1/9\,\sin(x/3) \implies M = max_{0\leq x\leq 1}|f^{(2)}(x)| = 1/9\,\sin(1/3)$. Sendo $a = 0$ e $b = 1$, N o número de subintervalos e $h = (b-a)/N = 1/N$, conclui-se da fórmula de erro da regra dos trapézios que

$$|I(f) - T_N(f)| \leq \frac{(b-a)^3\, M}{12\, N^2}.$$

Para $\epsilon = 10^{-4}$, pretende-se determinar o primeiro número natural N satisfazendo a desigualdade,

$$\frac{M}{12\,N^2} \leq \epsilon \ .$$

Por conseguinte, $N \geq \sqrt{\dfrac{M}{12\,\epsilon}} = \sqrt{\dfrac{\sin(1/3)\,10^4}{9 \times 12}} \simeq 5.50$. Assim, $N = 6$ e o número de nós a utilizar é $N + 1 = 7$.

4(a) Para $f(t, y) = 3\,t$, $t_0 = 0$, $t_i = t_0 + i\,h = i\,h$, tem-se a equação às diferenças

$$\begin{aligned} &y_{i+1} = y_i + h\,f(t_i, y_i) = y_i + \frac{3}{N^2}\,i, \qquad i = 0, 1, 2, \ldots \\ &\text{onde} \quad y_0 = -1 \ . \end{aligned}$$

4(b) A partir da equação às diferenças, resulta

$$\begin{aligned} y_1 &= -1 \\ y_2 &= y_1 + \frac{3}{N^2} = -1 + \frac{3}{N^2} \\ y_3 &= y_2 + \frac{3}{N^2} \times 2 = -1 + \frac{3}{N^2}\,(1 + 2) \\ y_4 &= y_3 + \frac{3}{N^2} \times 3 = -1 + \frac{3}{N^2}\,(1 + 2 + 3) \end{aligned}$$

Conclui-se facilmente por indução que

$$y_j = -1 + \frac{3}{N^2}\,(1 + 2 + \cdots + (j-1))\,, \qquad j = 1, 2, \ldots$$

Assim,

$$\begin{aligned} y_N \ &= -1 + \frac{3}{N^2}\,(1 + 2 + \cdots + (N-1)) \\ &= -1 + \frac{3}{N^2}\,\left(\frac{N}{2}\,(N-1)\right) = -1 + \frac{3}{2\,N}\,(N-1) \\ &= \frac{1}{2} - \frac{3}{2\,N} \ . \end{aligned}$$

Dado que $y(t) = -1 + 3\,t^2/2$, tem-se $y(1) = 1/2$. Consequentemente,

$$y(1) - y_N = \frac{3}{2\,N} \quad \Longrightarrow \lim_{N \to \infty} |y(1) - y_N| = 0 \ .$$

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & sons, New York, 1978.

[2] R. Bagnara, *A unified proof for the convergence of Jacobi and Gauss-Seidel methods*, SIAM Rev. 37, No. 1, 93-97, 1995.

[3] J.-P. Berrut, L. N. Trefethen, *Barycentric Lagrange interpolation*, SIAM Rev., 46(3), 501-517, 2004.

[4] J.-P. Berrut, Fascinante interpolation, *Bull. Soc. Frib. Sc. Nat.*, 83(1/2), 3-20, 1994.

[5] G. Birkhoff and G. Rota, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1978.

[6] W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, John Wiley & Sons, New York, 1992.

[7] J. P. Boyd, Finding the zeros of a univariate equation: Proxy root finders, Chebyshev interpolation, and the companion matrix, SIAM Rev., 55(2), 375-396, 2013.

[8] J. P. Boyd, *Solving transcendental equations, the Chebyshev Polynomial Proxy and other numerical root finders, perturbation series, and oracles*, SIAM, Philadelphia, 2014.

[9] M. Braun, *Differential Equations and Their Applications*, Springer, New York, 1993.

[10] R. L. Burden, J. Douglas Faires, *Numerical Analysis*, PWS-KENT, Boston, 1985.

[11] S. C. Chapra , R. P. Canale, *Métodos Numéricos para Engenharia*, McGraw–Hill, São Paulo, 2008.

[12] G. Dahlquist and A. Björck, *Numerical Methods in Scientific Computing*, Vol. I, SIAM, Philadelphia, 2008.

[13] J. F. Epperson, *On the Runge Example*, 1987, http://www.maa.org/sites/default/files/images/upload_library/22/Ford/Epperson329-341.pdf.

[14] J. Campos Ferreira *Introdução à Análise Matemática*, Fundação Calouste Gulbenkian, Lisboa, 1987.

[15] J. Harrison, *Decimal transcendentals via binary*, Computer Arithmetic, IEEE, 187-194, 2009.

[16] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & sons, New York, 1966.

[17] A. Gil, J. Segura, and N. Temme, *Numerical Methods for Special Functions*, Ch. 3, SIAM, Philadelphia, 2007,

http://www.siam.org/books/ot99/OT99SampleChapter.pdf.

[18] G. H. Golub and C. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, 1996.

[19] M. M. Graça and E. Sousa-Dias, *A unified framework for the computation of polynomial quadrature weights and errors*, arXiv:1203.4795v1, Mar 2012.

[20] M. M. Graça e P. T. Lima, *Matemática Experimental*, IST Press, 2007.

[21] M. M. Graça, *Maps for global separation of roots*, Electronic Transactions on Numerical Analysis (ETNA), Vol. 45, 241-256, 2016.

[22] J F. Grcar, *Mathematicians of Gaussian Elimination*, Notices of the AMS, Vol. 58, 6, 2011.

[23] A. Knoebel, R. Laubenbacher, J. Lodder, D. Pengelley *Mathematical Masterpieces, Further Chronicles by the Explorers*, Springer, 2007.

[24] R. Kress, *Numerical Analysis*, Springer, New York, 1998.

[25] P. Lima, *Métodos Numéricos da Álgebra*, disponível em

www.math.ist.utl.pt/plima/LMAC/mna.pdf .

[26] N. Madden, *John Todd and the development of modern Numerical Analysis*, Irish Math. Soc. Bulletin, 69, 11-23, 2012,

http://www.maths.tcd.ie/pub/ims/bull69/Madden.pdf.

[27] Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.

[28] J. M. Ortega, *Numerical Analysis – A second course*, Academic Press, New York, 1972.

[29] A. Ostrowski, *Lições de Cálculo Diferencial e Integral*, Vol. I, Fundação Calouste Gulbenkian, 1976, Lisboa.

[30] George M. Phillips, *Interpolation and Approximation by Polynomials*, Canadian Mathematical Society, Springer, New York, 2003.

[31] H. Pina, *Métodos Numéricos*, Escolar Editora, 2010.

[32] Z. Rached, *Arbitrary Order Iterations*, European Int. J. Science and Technology, Vol 2, 5, 191-195, 2013.

[33] D. A. Sanchez, R. C. Allen Jr., and W. T. Kyner, *Differential Equations*, Addison-Wesley, Massachusetts, 1988.

[34] J. Stillwell, *Elements of Algebra, Geometry, Numbers, Equations*, Springer, New York, 2001.

[35] J. Todd, *Basic Numerical Mathematics*, Vol. 2, Birkhäuser, Basel, 1977.

[36] J. Verbeke, R. Cools, *The Newton-Raphson method*, Int. J. Math. Educ. Sc. Tech. 26:2, 177-193, 2006.

[37] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

[38] S. Wolfram, *The Mathematica Book*, Wolfram Media, 2003.

[39] P. E. Zadunaisky, On the estimation of errors propagated in the numerical integration of ordinary differential equations, *Numer. Math.*, 27, 21-39 (1976).