Matemática Aplicada e Análise Numérica

Uma Introdução com Octave

Pedro Serranho

Secção de Matemática
Departamento de Ciências e Tecnologia
Universidade Aberta

2017

Conteúdo

Capítulo 1

Motivação

Um dos maiores desafios dos dias de hoje em ciência é a capacidade de fazer previsões. Este é um processo complicado com várias fases e que raramente passa por executar a experiência cujo resultado queremos descobrir, devido aos elevados custos ou pouco tempo disponível. Uma das possibilidades para conseguir prever passa por fazer a modelação do problema. Neste âmbito, a Matemática tem aplicações em várias áreas, desde as mais recentes tecnologias nas várias engenharias a áreas da saúde, passando pela meteorologia, arquitetura ou arte. A própria natureza apresenta padrões matemáticos que se repetem e podem portanto ser previstos.

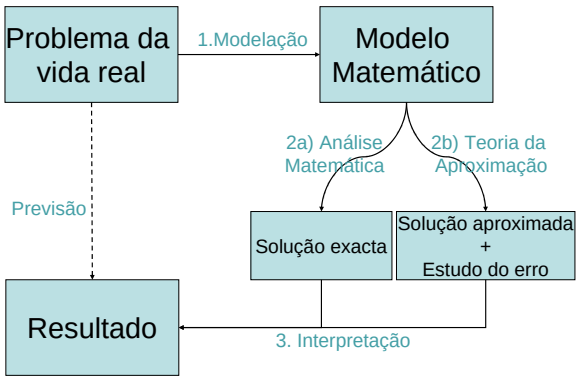


Figura 1.1: *Esquema do processo de previsão.*

Dado um problema da vida real, o primeiro passo é construir um modelo matemático que aproxima o problema. Geralmente, o modelo matemático implica

uma aproximação do problema inicial. O controlo do erro efetuado neste passo sai fora do âmbito desta unidade curricular. Tendo um modelo matemático é necessário resolvê-lo. A resolução poderá implicar resolver equações, inequações, maximizar ou minimizar funções, entre outros. Muitas vezes não é possível encontrar soluções exatas dos modelos, pelo que é necessário recorrer a métodos numéricos que aproximam as soluções com uma margem de erro controlável. Por último, dada a solução do modelo matemático é necessário interpretá-la para conseguir encontrar o resultado previsto. O processo de previsão precisa ainda de ser validado pelas observações da realidade para se mostrar fiável.

Neste texto vamos estar principalmente interessados em resolver modelos matemáticos (focando-nos na sua resolução numérica, isto é, em determinar aproximações da solução exata) e na interpretação da sua solução. Nos próximos capítulos vamos estudar alguns modelos e métodos simples, com aplicações várias nas áreas da engenharia e indústria, nomeadamente em casos em que a resolução analítica não é conhecida ou é demasiado morosa.

Começamos por apresentar um modelo simples, com base numa equação diferencial.

1.1 Exemplo conhecido: O modelo de queda livre

Um dos exemplos mais simples é o que modela a queda livre de um corpo. Vamos supor que um corpo em movimento vertical se encontra inicialmente a uma altura y_0 e velocidade v_0. Seja $y = y(t)$ a altura do corpo no instante t. Por definição, a velocidade $v = v(t)$ do corpo no instante t é a derivada da função posição $y(t)$, ou seja, $v(t) = y'(t)$. Sabemos ainda que a aceleração do corpo é a aceleração gravítica $g \approx 9.8$ no sentido negativo, pelo que a segunda derivada da função posição y é dada por

$$y''(t) = -g.$$

Se adicionarmos as condições iniciais, queremos encontrar a solução do problema

$$\begin{cases} y''(t) = -g, \quad t \geq 0 \\ y'(0) = v_0 \\ y(0) = y_0 \end{cases}$$

que modela a queda livre de um corpo com altura inicial y_0 e velocidade inicial v_0. Note-se que neste modelo estamos a desprezar as forças de atrito do ar, ilustrando que o modelo já é em si uma aproximação.

A solução da equação pode ser dada por integração direta, nomeadamente

$$y'(t) = \int y''(t)\,dt = -\int g\,dt = -g\,t + C$$

para algum $C \in \mathbb{R}$. Pela condição inicial para a velocidade $y'(0) = v_0$, temos $C = v_0$ pelo que

$$y'(t) = v_0 - g\,t.$$

Integrando de novo em t, obtemos

$$y(t) = \int y'(t)\,dt = \int v_0 - g\,t\,dt = v_0 t - g\frac{t^2}{2} + C$$

para algum $C \in \mathbb{R}$. De novo, pela condição inicial para a altura $y(0) = y_0$, temos $C = y_0$ pelo que a solução do problema é dada por

$$y(t) = y_0 + v_0 t - g\frac{t^2}{2}, \quad t \geq 0$$

o que nos dá a equação (já conhecida da Física) do movimento de uma partícula uniformemente acelerada (com aceleração $-g$) em movimento retilíneo.

Este exemplo ilustra um caso simples de resolução analítica de um modelo para um fenómeno físico. Outros casos particulares de equações diferenciais podem ser também resolvidos analiticamente, como por exemplo equações diferenciais ordinárias lineares, ou de coeficientes contantes, ou de variáveis separáveis. Para uma análise mais cuidada destes e de outros casos de possibilidades de resolução analítica, pode consultar por exemplo [1, 2, 3].

No entanto é fácil de compreender que em casos mais complexos o processo para obter uma solução analítica pode ser bem mais complicado ou mesmo não ser possível. Por exemplo se for introduzido no modelo a resistência do ar em forma de força de atrito F_a, teríamos pela segunda lei de Newton ($F = m.a$) que a aceleração do corpo seria dada por

$$y''(t) = -g + \frac{F_a}{m}$$

em que m é a massa do corpo e a força de atrito é geralmente considerada como proporcional à velocidade

$$F_a = Ay'(t),$$

em que o termo A depende da forma do objeto e da densidade do ar, entre outros. Em particular, se pensarmos que a densidade do ar depende da altura (e para facilitar) de forma linear, isto é, $A = Cy(t)$ ficamos com a equação

$$y''(t) = -g + C\frac{y(t)\,y'(t)}{m}$$

que deixa de ser uma equação linear na solução y.

Assim é necessário encontrar formas expeditas de obter soluções de problemas mais complexos motivados por aplicações reais, por exemplo que envolvam equações não-lineares ou equações às derivadas parciais, uma vez que na maioria dos problemas com aplicações práticas a solução analítica não é possível de obter.

Conforme já foi referido nesta introdução, geralmente o preço a pagar por ter um método para resolver estas equações é que a solução encontrada deixa de ser exata para ser apenas aproximada. É neste contexto que avançamos para os próximos capítulos, começando evidentemente por estabelecer formas apropriadas de medir o erro cometido numa aproximação.

Capítulo 2

Teoria do erro

Associado a qualquer tipo de medição está sempre o erro cometido, partindo do princípio que na prática uma mediação nunca é exata. Dependendo da escala do aparelho de medida e da experiência do técnico de medição, podemos obter um erro mais ou menos grosseiro. Existem várias formas de medir o erro. A mais usual é o chamado **erro absoluto** $e_{\tilde{x}}$ da aproximação $\tilde{x}$, em que consideramos o módulo da diferença entre o valor exato x e a aproximação $\tilde{x}$, isto é,

$$e_{\tilde{x}} = |x - \tilde{x}|.$$

No caso geral, em que x não é um valor real (pode ser um vector, uma matriz, uma função,...) convém considerar uma métrica adequada, substituindo-se o módulo por uma norma adequada, isto é,

$$e_{\tilde{x}} = ||x - \tilde{x}||.$$

No entanto o erro absoluto nem sempre é um bom indicador da qualidade da aproximação. Por exemplo, cometer um erro absoluto de um metro na distância entre Lisboa e Porto pode ser considerado bom, mas cometer o mesmo erro absoluto na medição da altura de uma pessoa é considerado um erro grosseiro. Desta forma, considera-se também o **erro relativo** $\delta_{\tilde{x}}$ dado por

$$\delta_{\tilde{x}} = \frac{||x - \tilde{x}||}{||x||}$$

que nos indica a percentagem de erro cometido em relação à grandeza do valor a aproximar. Temos então a seguinte definição.

Definição 2.1 (Erro absoluto e relativo)**.**

Seja $\tilde{x}$ a aproximação de um valor exato x e seja $||.||$ uma norma adequada para x.

Chama-se **erro absoluto** $e_{\tilde{x}}$ a

$$e_{\tilde{x}} = ||x - \tilde{x}|| \tag{2.1}$$

e **erro relativo** $\delta_{\tilde{x}}$ a

$$\delta_{\tilde{x}} = \frac{||x - \tilde{x}||}{||x||}. \tag{2.2}$$

O erro relativo indica a percentagem de erro em relação à grandeza de x.

Exercício 2.2. Determine os erros relativos e absolutos associados às aproximações seguintes:

(a) Um lápis mede exatamente $12.23cm$. Com a ajuda de uma régua com escala até aos milímetros, obteve-se a medição $12.2cm$.

(b) Conseguiu-se mostrar teoricamente que existem 64.345×10^{23} moléculas de determinada substância num recipiente. No entanto, utilizando métodos laboratoriais apropriados obteve-se o valor experimental de 64.554×10^{23} moléculas.

(c) De determinado ponto da terra, sabe-se que uma estrela conhecida está na direção $(0.23, 0.65, 0.72)$. No entanto, o telescópio utilizado para avistar a estrela foi apontado na direção $(0.2, 0.6, 0.7)$. (Nota: utilize a norma usual para vectores.)

Resposta.

(a) O erro absoluto é $e_{\tilde{x}} = 0.03cm$ enquanto que o erro relativo é aproximadamente 0.00245.

(b) O erro absoluto é $e_{\tilde{x}} = 2.09 \times 10^{22}$ moléculas enquanto que o erro relativo é aproximadamente 0.00325.

(c) O erro absoluto é aproximadamente $e_{\tilde{x}} = 0.0616$ enquanto que o erro relativo é aproximadamente 0.0618.

2.1 Erros de Arredondamento

O arredondamento é outro fator que gera erros. Consideremos um número num sistema de vírgula flutuante

$$x = \pm 0.a_1 a_2 \dots a_n \cdots \times 10^t$$

em que $a_j,\ j \in \mathbb{N}$ são os algarismos que compõe o número e $t \in \mathbb{N}$ é o expoente. Aos algarismos não afetados por erros chamam-se **algarismos significativos**. Em particular, se os dados recolhidos são apresnetados com n algarismos significativos, após operações aritméticas o resultado deve ser apresentado com o mesmo número de algarismos significativos, pois apenas esses são representativos.

Note-se também que sempre que usamos uma representação de um número para cálculo, a sua representação é necessariamente finita e limitada pela capacidade de memória da máquina. Desta forma, seja por restrições de memória ou por restrições de precis ao, é sempre necessário considerar o arredondamento do número x para uma aproximação da forma

$$x = \pm 0.a_1 a_2 \dots a_n \times 10^t$$

com $a_j \in \{0, 1, 2, \dots, 9\}$ e $t \in [-b, b] \cap \mathbb{N}$.

Podemos então considerar dois tipos de arredondamento. No **arredondamento por corte** ao n-ésimo algarismo tomamos a aproximação

$$\tilde{x} = \pm 0.a_1 a_2 \dots a_n \times 10^t.$$

No caso de **arredondamento simétrico** tomamos a aproximação

$$\tilde{x} = \pm 0.a'_1 a'_2 \dots a'_n \times 10^t.$$

em que os algarismos $a'_j,\ j = 1, 2, \dots n$ resultam do arredondamento por corte da soma de x com $0.5 \times 10^{t-n}$. Este processo consiste simplesmente em arredondar x por defeito se $a_{n+1} \in \{0, 1, \dots, 4\}$ e por excesso se $a_{n+1} \in \{5, 6, \dots, 9\}$.

Nota 2.3. Salvo nota em contrário, consideramos neste texto arredondamento simétrico, uma vez que em geral o erro cometido é menor.

Exercício 2.4. Determine os arredondamentos por corte e simétrico ao 3º algarismo de:

(a) $x = 12.3454$;

(b) $x = 0.0001255$;

(c) $x = 9.298$.

Resposta.

(a) Tanto por corte como por arredondamento simétrico obtemos $\tilde{x} = 12.3$;

(b) Por corte obtemos $\tilde{x} = 0.125 \times 10^{-3}$ enquanto que por arredondamento simétrico obtemos $\tilde{x} = 0.126 \times 10^{-3}$;

(c) Por corte obtemos $\tilde{x} = 9.29$ enquanto que por arredondamento simétrico obtemos $\tilde{x} = 9.30$;

2.2 Propagação de erros

O estudo da propagação de erro com aplicação de funções e operações a valores aproximados é também fundamental no contexto da aproximação numérica e na análise numérica. Assim, pretende-se compreender como se propaga o erro de um valor aproximado quando lhe aplicamos uma função ou o usamos para somar ou multiplicar por outro valor.

Para isso precisamos da fórmula de Taylor, que será introduzida em detalhe no secção 5.1. Para já, e como estamos num capítulo introdutório, utilizemos a aproximação da derivada de uma função real f diferenciável, dada por

$$f'(x) \approx \frac{f(\tilde{x}) - f(x)}{\tilde{x} - x} \tag{2.3}$$

quando $\tilde{x} \approx x$, como se assume no caso de $\tilde{x}$ ser uma aproximação[1] de x. Desta forma temos

$$f(\tilde{x}) - f(x) \approx f'(x)(\tilde{x} - x).$$

Assim temos a relação entre os erros absolutos de $\tilde{x}$ e $f(\tilde{x})$ dada por

$$e_{f(\tilde{x})} \approx |f'(x)| e_{\tilde{x}}. \tag{2.4}$$

Exercício 2.5. Determine a forma como se propaga o erro absoluto com as seguintes funções:

(a) $f(x) = x^n$ com $x \neq 0$, em particular para $x = 1$;

(b) $f(x) = e^x$, em particular para $x = 0$;

[1] O simbolo $\approx$ significa "é aproximadamente igual a".

Resposta.

(a) $e_{f(\tilde{x})} = n|x|^{n-1}e_{\tilde{x}}$, em particular para $x = 1$ temos $e_{f(\tilde{x})} \approx ne_{\tilde{x}}$.

(b) $e_{f(\tilde{x})} = e^x e_{\tilde{x}}$, em particular para $x = 0$ temos $e_{f(\tilde{x})} \approx e_{\tilde{x}}$.

De forma semelhante, a partir da Fórmula de Taylor ou da aproximação do plano tangente ao gráfico de uma função real e diferenciável de duas variáveis reais, temos

$$f(\tilde{x}, \tilde{y}) - f(x, y) \approx \frac{\partial f}{\partial x}(x, y)(\tilde{x} - x) + \frac{\partial f}{\partial y}(x, y)(\tilde{y} - y),$$

ou seja, temos a aproximação para a propagação do erro absoluto dada por

$$e_{f(\tilde{x},\tilde{y})} \approx \left|\frac{\partial f}{\partial x}(x, y)\right| e_{\tilde{x}} + \left|\frac{\partial f}{\partial y}(x, y)\right| e_{\tilde{y}}. \tag{2.5}$$

Como já referimos, o erro absoluto não é a melhor forma de medir a qualidade das aproximações, uma vez que não reflete a grandeza do valor a aproximar. Desta forma, é importante estudar a propagação de erros relativos com operações. O resultado é apresentado no teorema seguinte, sendo que a prova sai diretamente das expressões (2.4) e (2.5).

Teorema 2.6 (Propagação de erros)**.**
Seja $f : \mathbb{R} \to \mathbb{R}$ diferenciável. Então

$$\delta_{f(\tilde{x})} \approx \left|\frac{x\, f'(x)}{f(x)}\right| \delta_{\tilde{x}}. \tag{2.6}$$

Seja $f : \mathbb{R}^2 \to \mathbb{R}$ diferenciável. Então

$$\delta_{f(\tilde{x},\tilde{y})} \approx \left|\frac{x}{f(x, y)}\frac{\partial f}{\partial x}(x, y)\right| \delta_{\tilde{x}} + \left|\frac{y}{f(x, y)}\frac{\partial f}{\partial y}(x, y)\right| \delta_{\tilde{y}}. \tag{2.7}$$

Exercício 2.7. Mostre que a propagação do erro relativo na multiplicação é dada por

$$\delta_{\widetilde{x\tilde{y}}} \approx \delta_{\tilde{x}} + \delta_{\tilde{y}}.$$

e que a propagação do erro no caso da soma é dada por

$$\delta_{\widetilde{x+y}} \approx \left|\frac{x}{x + y}\right| \delta_{\tilde{x}} + \left|\frac{y}{x + y}\right| \delta_{\tilde{y}}.$$

O resultado do exercício anterior mostra que a propagação de erros com a multiplicação é moderada, uma vez que é a soma dos erros relativos das aproximações de cada um dos elementos do produto. Por outro lado, a propagação de erros no caso da soma pode ser catastrófica no caso de $x \approx -y$. Neste caso, a propagação do erro inicial de cada uma das parcelas pode ser muito amplificada. A este fenómeno dá-se o nome de **cancelamento subtrativo**. Note-se que a soma no caso $x \approx -y$ é equivalente à diferença entre dois valores aproximados muito próximos, daí a designação de cancelamento subtrativo.

Exercício 2.8 (Exemplo de Cancelamento Subtrativo). Considere a função dada por

$$f(x) = \frac{1 - \cos^2 x}{\sin^2(x)}, \quad x \neq 0;$$

(a) Mostre que $f(x) = 1$ no seu domínio, recorrendo à fórmula fundamental da trigonometria.

(b) Mostre em particular que

$$\lim_{x \to 0} f(x) = 1.$$

(c) Recorrendo ao Octave, calcule pela expressão indicada $f(10^{-n})$, para os valores de $n \in \{5, 6, 7, 8, 9\}$.

(d) Justifique o resultado obtido.

Resolução.

(a) Pela fórmula fundamental da trigonometria, temos

$$\cos^2 x + \sin^2(x) = 1,$$

logo $\sin^2(x) = 1 - \cos^2(x)$, pelo que $f(x) = 1$ no seu domínio.

(b) Trivial, pela alínea anterior.

(c) Temos

$$f(10^{-5}) = 1.0000; f(10^{-6}) = 1.0001; f(10^{-7}) = 0.99920; f(10^{-8}) = f(10^{-9}) = 0;$$

(d) O resultado obtido é inesperado, no sentido em que o resultado do cálculo numérico em Octave parece contradizer a alínea b). Isto deve-se a um caso de cancelamento subtrativo, uma vez que $\cos^2(x) \approx 1$ quando $x \approx 0$. Assim, para valores de x muito próximos de zero, a precisão da máquina arredonda $1 - \cos^2(x)$ por zero.

2.3 Outras Formas de Erro

Além do erro de arredondamento e da sua propagação por aplicação de funções, existem ainda outras formas de erro. No âmbito deste texto o mais relevante será o **erro de truncatura** . Este erro deve-se em particular a truncar uma soma de parcelas (finitas ou infinitas) para um número finito inferior de parcelas. Por exemplo, suponhamos que pretendemos aproximar o valor da soma da série

$$S = \sum_{n=1}^{\infty} \frac{\cos(n)}{2^n}.$$

Em particular sabemos que

$$\left|\frac{\cos(n)}{2^n}\right| \leq \frac{1}{2^n},$$

pelo que a série é absolutamente convergente e logo convergente. Assim, podemos aproximar a soma da série S por

$$S \approx S_N = \sum_{n=1}^{N} \frac{\cos(n)}{2^n},$$

para algum $N \in \mathbb{N}$ finito. Assim, estamos a introduzir um erro de truncatura dado por

$$S - S_N = \sum_{n=N+1}^{\infty} \frac{\cos(n)}{2^n},$$

Em particular sabemos pela soma de uma série geométrica que o erro absoluto de truncatura é neste caso majorado por

$$|S - S_N| \leq \sum_{n=N+1}^{\infty} \left|\frac{\cos(n)}{2^n}\right| \leq \sum_{n=N+1}^{\infty} \frac{1}{2^n} = \frac{1}{2^{N+1}} \underbrace{\sum_{n=0}^{\infty} \frac{1}{2^n}}_{=\frac{1}{1-1/2}} = \frac{1}{2^N}.$$

Como exemplo se truncarmos a soma e a reduzirmos apenas a $N = 3$ parcelas, temos a aproximação

$$S \approx S_3 = \frac{\cos(1)}{2} + \frac{\cos(2)}{4} + \frac{\cos(3)}{8} = 0.042365$$

com majorante de erro $|S - S_3| \leq 0.125$. Por outro lado, se truncarmos a série a partir da décima parcela, temos

$$S \approx S_{10} = \frac{\cos(1)}{2} + \frac{\cos(2)}{4} + \frac{\cos(3)}{8} + ... + \frac{\cos(10)}{2^{10}} = 0.028102$$

com majorante de erro $|S - S_{10}| \leq 0.00097656$.

Voltaremos ao erro de truncatura no Capítulo 5, em particular na nota 5.4 sobre o erro de truncatura da fórmula de Taylor.

Capítulo 3

Princípios Básicos de Análise Funcional

Neste capítulo vamos recordar alguns resultados de análise funcional que serão necessários para os resultados dos capítulos seguintes. Na maior parte dos casos não ilustraremos as provas, remetendo para qualquer livro de Análise Funcional para o efeito (ver por exemplo [4]). Em particular, alguns livros de análise numérica (ver por exemplo [5]), também contêm as demonstrações dos resultados neste capítulo.

3.1 Espaços normados e de Banach

Comecemos então com a definição de norma, que será a base para a definição de um espaço normado e posteriormente de Banach.

Definição 3.1 (Norma)**.**
Seja X um espaço linear real (ou complexo). A função $\|.\| : X \to \mathbb{R}$ diz-se uma norma se para todo $x, y \in X$ e $\alpha \in \mathbb{R}$ (ou $\alpha \in \mathbb{C}$) se verificarem as seguintes condições

(a) Positividade: $\|x\| \geq 0$;

(b) Definitividade: $\|x\| = 0 \Leftrightarrow x = 0$;

(c) Homogeneidade: $\|\alpha x\| = |\alpha| \, \|x\|$;

(d) Desigualdade Triangular : $\|x + y\| \leq \|x\| + \|y\|$;

Definição 3.2 (Espaço Normado, Completo, de Banach)**.**

Um espaço linear X real (ou complexo) diz-se **normado**, se estiver equipado com uma norma.

Um espaço normado X diz-se **completo** se qualquer sucessão de Cauchy$(x_n)_{n\in\mathbb{N}}$ de elementos em X, isto é,

$$\forall_{\varepsilon>0}\,\exists_{N\in\mathbb{N}}\,\forall_{n,m>N}\,\|x_n - x_m\| < \varepsilon,$$

é convergente para um elemento $x \in X$.

A um espaço normado completo chama-se espaço de **Banach**.

Exemplo 3.3 (Espaços de Banach). Alguns exemplos de espaços de Banach são os seguintes:

(a) Os espaços $\mathbb{R}^n$ e $\mathbb{C}^n$ são espaços de Banach com as normas dadas por

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

com $p \in \mathbb{N}$.

(b) Os espaços $\mathbb{R}^n$ e $\mathbb{C}^n$ são espaços de Banach com a norma do máximo dada por

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|.$$

(c) Os espaços de sucessões reais ou complexas $(x_n)_{n\in\mathbb{N}}$ dado por

$$\ell^p = \left\{(x_n)_{n\in\mathbb{N}} : \|x\|_p < \infty\right\}$$

com a norma

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{\infty} |x_i|^p}$$

são espaços de Banach.

(d) O espaço de sucessões reais ou complexas $(x_n)_{n\in\mathbb{N}}$ dados por

$$\ell^\infty = \left\{(x_n)_{n\in\mathbb{N}} : \|x\|_\infty < \infty\right\}$$

com a norma do máximo

$$\|x\|_\infty = \max_{i\in\mathbb{N}} |x_i|.$$

é um espaço de Banach.

(e) Os espaços de funções

$$L^p = \{f : [a,b] \to \mathbb{R} : \|f\|_p < \infty\}$$

com a norma

$$\|x\|_p = \sqrt[p]{\int_a^b |f(x)|^p dx}$$

são espaços de Banach.

Precisamos também de recordar alguns resultados sobre operadores sobre espaço normados. Começamos por definir um operador linear.

Definição 3.4 (Operador linear)**.**
Seja $A : X \to Y$ um operador do espaço X para o espaço Y real (ou complexo). O operador A diz-se **linear** se

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay, \tag{3.1}$$

para qualquer $x, y \in X$ e $\alpha, \beta \in \mathbb{R}$ (ou $\mathbb{C}$).

Definimos também a norma de um operador induzida pelo domínio e contra-domínio, da seguinte forma.

Definição 3.5 (Norma de um operador, Operador limitado)**.**
Sejam X, Y dois espaços lineares normados. Define-se como **norma do operador** linear $A : X \to Y$ (induzida pelas normas $\|.\|_X$ e $\|.\|_Y$ em X e Y, respetivamente) por

$$\|A\| := \sup_{\|x\|_X = 1} \|Ax\|_Y. \tag{3.2}$$

Se $\|A\| < \infty$ diz-se que o operador A é **limitado**.

Por facilidade de notação, nos exemplos seguintes vamos deixar cair o índice X e Y das normas respetivas. Temos o seguinte majorante para operadores limitados.

Teorema 3.6.
Nas condições da definição anterior, se A é limitado temos em particular que

$$\|Ax\| \leq \|A\| \, \|x\|, \quad \forall x \in X. \tag{3.3}$$

De forma semelhante, se A e B são operadores limitados, então

$$\|ABx\| \leq \|A\| \, \|B\| \, \|x\|, \quad \forall x \in X, \tag{3.4}$$

ou seja, $\|AB\| \leq \|A\| \, \|B\|$.

Demonstração. A desigualdade (3.3) mostra-se simplesmente por

$$\|Ax\| = \left\| A\left(\frac{x}{\|x\|}\right)\right\| \|x\| \leq \sup_{\|x\|=1} \|Ax\| \, \|x\| = \|A\| \, \|x\| \, .$$

A demonstração de (3.4) é agora trivial, considerando $y = Bx$ e aplicando duas vezes a desigualdade (3.3). □

Outra definição importante é a de operador contínuo.

Definição 3.7 (Operador contínuo)**.**
Sejam X, Y dois espaços lineares normados. O operador $A : X \to Y$ diz-se **contínuo no ponto** x se para qualquer sucessão $(x_n)_{n\in\mathbb{N}}$ convergente para x

$$x_n \to x$$

temos $Ax_n \to Ax$. Se A é contínuo para todo o $x \in X$, o operador diz-se **contínuo em** X.

Para operadores lineares, temos a seguinte equivalência entre continuidade num ponto e continuidade global.

Teorema 3.8.
Sejam X, Y dois espaços lineares normados e $A : X \to Y$ um operador linear. Então A é contínuo num ponto $x_0 \in X$ se e só se é contínuo em X.

Demonstração. Seja A contínuo em x_0 e seja $x_n \to x \neq x_0$. Então para $y_n := x_n - x + x_0$ temos $y_n \to x_0$ e logo $Ay_n \to Ax_0$. Assim, por linearidade do operador, temos

$$Ax_n - Ax = Ay_n - Ax_0 \to 0.$$

logo $Ax_n \to Ax$. Como x é arbitrário, o resultado está demonstrado. □

Temos também para operadores lineares a equivalência entre operador contínuo e limitado.

Teorema 3.9.
Sejam X, Y dois espaços lineares normados e $A : X \to Y$ um operador linear. Então A é contínuo se e só se é limitado.

Demonstração. Seja A limitado e $x_n \to 0$. Então

$$\|Ax_n\| \leq \|A\| \, \|x_n\| \to 0$$

logo $Ax_n \to 0$ e pelo teorema 3.8 temos que A é contínuo.

Seja A contínuo e assumamos por absurdo que não é limitado, ou seja, existe uma sucessão $(x_n)_{n\in\mathbb{N}}$ tal que $\|x_n\| = 1$ e $\|A\| > n$. Definindo $y_n = x_n/\|Ax_n\|$, temos por linearidade do operador que $\|Ay_n\| = 1$. Por outro lado

$$\|y_n\| \leq \frac{1}{n} \to 0$$

pelo que por continuidade do operador temos $Ay_n \to 0$, o que conclui a demonstração por absurdo. □

Definição 3.10 (Operador invertível)**.**

Seja $A : X \to Y$ um operador bijetivo, isto é, para cada $b \in Y$ existe um único $x \in X$ tal que

$$Ax = b.$$

Chama-se ao operador $A^{-1} : Y \to X$ tal que

$$Ax = y \Leftrightarrow A^{-1}b = x,$$

o operador **inverso** de A.

Para o caso de matrizes, sabemos da álgebra linear [6] que uma matriz A é invertível se e só se

$$\min\{|\lambda| : \lambda \text{ é valor próprio de A}\} > 0,$$

ou por outras palavras se o seu determinante (dado pelo produto dos valores próprios) for diferente de zero.

Para o caso de operadores em espaços de Banach, existem vários resultados que garantem a invertibilidade de um operador. Um deles é o seguinte.

Teorema 3.11.

Seja X um espaço de Banach e $B : X \to X$ um operador linear com norma $\|B\| < 1$. Seja $I : X \to X$ o operador identidade. Então o operador $I - B$ e invertível e a sua inversa dada por

$$(I - B)^{-1} = \sum_{n=0}^{\infty} B^n \tag{3.5}$$

tem norma majorada por

$$\|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Demonstração. Apresentamos apenas a ideia geral da prova, deixando os detalhes técnicos para [5]. Começamos por verificar que a série do membro direito de (3.5) é convergente, uma vez que $\|B\| < 1$ e da análise matemática [7] sabemos que se a série das normas $\sum_{n=1}^{\infty} \|B^n x\|$ converge então a série (3.5) também converge. Como

$$(I - B) \sum_{n=0}^{\infty} B^n = \sum_{n=0}^{\infty} B^n - \sum_{n=1}^{\infty} B^n = I$$

e

$$\sum_{n=0}^{\infty} B^n (I - B) = \sum_{n=0}^{\infty} B^n - \sum_{n=1}^{\infty} B^n = I,$$

mostramos que a expressão (3.5) representa uma inversa de $I - B$. A estimativa para a norma sai da majoração da série (3.5) pela série das normas e da expressão para a soma de uma série geométrica. □

3.1.1 Espaços de Banach de dimensão finita

A resolução numérica de problemas em matemática aplicada resume-se muitas vezes à resolução de sistemas lineares ou à utilização de matrizes para formulações de aproximações lineares do problema. Desta forma assume particular importância, o estudo de espaços matriciais e vetoriais. Nesta secção vamos recordar alguns resultados para espaços de dimensão finita, em que estes se incluem.

Começamos por recordar que qualquer espaço normado de dimensão finita é um espaço de Banach. Para isso, recordamos o Teorema de Bolzano-Weierstrass, que será a base para essa prova.

Teorema 3.12 (Bolzano-Weierstrass)**.**
Qualquer sucessão limitada $(x_n)_{n\in\mathbb{N}}$ em $\mathbb{R}^n$, isto é, tal que existe um $C > 0$ tal que

$$\|x_n\| < C, \forall n \in \mathbb{N},$$

tem uma subsucessão convergente , isto é, existe um $x \in \mathbb{R}^n$ tal que

$$x_{n(k)} \to x, \quad k \to \infty.$$

Demonstração. Remetemos para um livro de Análise Real, como por exemplo [8, 7]. □

O teorema de Bolzano-Weierstrass pode ser extrapolado para qualquer espaço normado de dimensão finita, pelo teorema seguinte.

Teorema 3.13.
Num espaço normado X de dimensão finita, qualquer sucessão $(x_n)_{n\in\mathbb{N}}$ limitada tem uma subsucessão convergente.

Demonstração. Seja $\{u_1, u_2, \ldots, u_k\}$ uma base de X. Então temos a representação

$$x_n = \sum_{i=1}^{k} \alpha_{n,i} u_i.$$

como a sucessão $(x_n)_{n\in\mathbb{N}}$ é limitada então cada uma das sucessões do vector de coeficientes $(\alpha_n)_{n\in\mathbb{N}}$ é também limitada em $\mathbb{R}^n$. Assim, pelo teorema de Bolzano-Weierstrass 3.12 existe $\alpha \in \mathbb{R}^n$ tal que

$$\alpha_{n(k)} \to \alpha$$

logo

$$x_{n(j)} \to x = \sum_{i=1}^{k} \alpha_i u_i.$$

□

Um outro resultado que advém do teorema de Bolzano-Weierstrass é o seguinte.

Teorema 3.14 (Equivalência de normas em espaços de dimensão finita)**.**
Num espaço normado X de dimensão finita, todas as normas são equivalentes, isto é, dadas duas normas $\|.\|_1$ e $\|.\|_2$ existem duas constantes c_1 e c_2 tal que

$$c_1\|x\|_1 \leq \|x\|_2 \leq c_1\|x\|_2 \forall x \in X.$$

Demonstração. Ver [5, Thm.3.8]. □

Estamos agora em condições de formular o resultado.

Teorema 3.15.
Um espaço normado X de dimensão finita é um espaço de Banach.

Demonstração. Queremos mostrar que qualquer sucessão de Cauchy $(x_n)_{n\in\mathbb{N}}$ é convergente. Seja $\{u_1, u_2, \ldots, u_k\}$ uma base de X e consideremos a representação da sucessão de Cauchy dada por

$$x_n = \sum_{i=1}^{k} \alpha_{n,i} u_i.$$

Pelo teorema 3.14 temos que existe $C > 0$ tal que

$$\max_{i=1,\ldots,k} |\alpha_{n,i} - \alpha_{m,i}| \leq C\|x_n - x_m\|, \forall n, m \in \mathbb{N}.$$

logo $(\alpha_{n,i})_{n\in\mathbb{N}}$ é uma sucessão de Cauchy em $\mathbb{C}$ (ou $\mathbb{R}$) para cada $i = 1, \ldots, k$,, logo é convergente (pois $\mathbb{C}$ (ou $\mathbb{R}$) é completo). Assim temos a convergência de x_n para

$$x_n \to x = \sum_{i=1}^{k} \alpha_i u_i.$$

em que α_i é o limite da sucessão $(\alpha_{n,i})_{n\in\mathbb{N}}$ para cada $i = 1, \ldots, k$,. □

3.2 Espaços pré-Hilbertianos e de Hilbert

Passemos então a recordar a definição de produto interno e espaço de Hilbert.

Definição 3.16 (Produto interno, Espaço Pré-Hilbertiano)**.**

Uma função $(.,.) : X \times X \to \mathbb{C}$ (ou $\mathbb{R}$) definida para um espaço linear X complexo (ou real) diz-se um **produto interno** se satisfizer as condições

(a) Positividade: $(x, x) \geq 0$;

(b) Definitividade: $(x, x) = 0 \Leftrightarrow x = 0$;

(c) Simetria: $(x, y) = \overline{(y, x)}$;

(d) Linearidade: $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$;

para qualquer $x, y, z \in X$ e $\alpha, \beta \in \mathbb{C}$ (ou $\mathbb{R}$).

Um espaço linear equipado com um produto interno diz-se **Pré-Hilbertiano** .

Relembramos de seguida a desigualdade de Cauchy-Schwarz, que será útil mais adiante.

Teorema 3.17 (Cauchy-Schwarz)**.**
Temos a desigualdade para um produto interno

$$|(x,y)|^2 \leq (x,x)(y,y). \tag{3.6}$$

Demonstração. Para $x = 0$ é trivial. Para $x \neq 0$, temos

$$0 \leq (\alpha x + \beta y, \alpha x + \beta y) = |\alpha|^2(x,x) + 2\mathrm{Re}\Big(\alpha\bar{\beta}(x,y)\Big) + |\beta|^2(y,y).$$

O resultado obtém-se escolhendo

$$\alpha = -\frac{\overline{(x,y)}}{\sqrt{(x,x)}}, \quad \beta = \sqrt{(x,x)}.$$

□

A desigualdade de Cauchy-Schwartz é a base para mostrar que é sempre possível definir uma norma num espaço Pré-Hilbertiano.

Teorema 3.18 (Norma induzida por produto interno)**.**
Um espaço pré-Hilbertiano é sempre um espaço normado com norma definida por

$$||x|| = \sqrt{(x,x)}.$$

Demonstração. Basta mostrar que a norma anterior satisfaz os axiomas da definição 3.1 de norma. Deixamos a prova como exercício, sugerindo a utilização da desigualdade de Cauchy-Schwartz para obter a desigualdade triangular. □

Note-se que utilizando a norma, a desigualdade de Cauchy-Schwartz (3.6) pode ser escrita como

$$|(x,y)| \leq \|x\| \, \|y\|. \tag{3.7}$$

Além disso o resultado 3.18 mostra que um espaço de Hilbert é sempre um espaço de Banach, pelo que os resultados válidos para espaços de Banach são válidos em espaços de Hilbert. Mais ainda, havendo a definição de norma, no contexto de espaços pré-Hilbertianos podemos também definir um espaço completo.

Definição 3.19 (Espaço de Hilbert)**.**
Um espaço Pré-Hilbertiano completo diz-se um espaço **Hilbertiano** ou **espaço de Hilbert**.

Capítulo 4

Condicionamento

Em matemática, antes de procurar uma solução de determinado problema, é boa prática estudar primeiro se a solução existe (de outra forma, não faz sentido procurá-la) e se é única (caso contrário, determinado método pode encontrar uma solução, mas essa pode ser uma solução que não interessa dentro do leque de soluções possíveis). Do ponto de vista da análise numérica, em que se procura uma solução aproximada do problema, existe ainda outro aspecto de grande importância: o bom-condicionamento. Um problema diz-se bem condicionado se a solução depender continuamente dos dados, isto é, se pequenos erros nos dados provocam pequenos erros no resultado. Este aspecto é fulcral para a obtenção de aproximações de soluções exatas. De facto, se o problema for mal-condicionado, uma resolução numérica descuidada pode levar a que pequenos erros origem uma aproximação muito distante da solução exata, o que não é de todo desejável. A importância do condicionamento na resolução de um problema é ainda mais evidente se considerarmos uma situação real, em que na maior parte dos casos os dados são obtidos por medição e são portanto afetados de erros.

Desta forma, Hadamard [9] sugeriu a seguinte definição para problema bem-posto.

Definição 4.1 (Problema Bem-Posto)**.**

Um problema diz-se **bem-posto** no sentido de Hadamard se verificar em simultâneo as condições de:

(a) Existência de solução;

(b) Unicidade de solução;

(c) Bom-condicionamento (dependência contínua dos dados), isto é, pequenos erros nos dados provocam pequenos erros nos resultados;

Caso o problema não satisfaça uma das condições acima diz-se **mal-posto**. Em particular, se um problema não verificar a condição (c) diz-se **mal-condicionado**.

Na prática, diz-se que o problema associado à aplicação do operador linear A dada por

$$Ax = b \tag{4.1}$$

é bem condicionado num ponto x, se existir uma constante $C \geq 0$ tal que

$$\delta_{\tilde{b}} \leq C\delta_x, \forall x \in V_x,$$

em que V_x representa uma vizinhança de x. Isto quer dizer que o erro relativo dos resultados é controlado pelo erro relativo nos dados, ou seja, que o resultado depende continuamente dos resultados.

Como introdução ao estudo de condicionamento de problemas lineares, temos de estabelecer uma norma para o operador linear A. Começamos pelo caso mais simples em que A tem dimensão finita, ou seja, pode ser representado por uma matriz.

4.1 Normas de matrizes

Nesta secção, salvo nota em contrário vamos considerar uma matriz complexa A de dimensões $n \times n$ com entradas $a_{i,j}$ para $i, j = 1, 2, \ldots, n$ e dois vectores coluna complexos $x, b \in \mathbb{C}^n$ dados por

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \ldots & a_{ij} & \ldots & a_{in} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nj} & \ldots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \tag{4.2}$$

Para prosseguirmos com a análise de condicionamento precisamos de definir normas de uma matriz. Uma forma de o fazer é considerar a matriz $n \times n$ como um operador no espaço linear dos vectores coluna $\mathbb{R}^n$ (ou $\mathbb{C}^n$). Dessa forma podemos considerar a norma induzida pela norma dos vectores, no espírito da definição de norma de operador (3.2). Temos então o seguinte teorema, que caracteriza algumas normas de matrizes, que nos serão úteis no decorrer deste texto.

Teorema 4.2 (Normas de Matrizes)**.**

A normas no espaço de vectores coluna $\mathbb{R}^n$ (ou $\mathbb{C}^n$) dadas por

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{4.3}$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \tag{4.4}$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \tag{4.5}$$

induzem (no sentido da definição 3.5) as normas no espaço linear de matrizes reais ou complexas de dimensão $n \times n$ dadas por

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \tag{4.6}$$

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{4.7}$$

$$\|A\|_2 = \sqrt{\rho\left(A^* A\right)} \tag{4.8}$$

respetivamente, onde A^* é a matriz transposta (conjugada) de A e o ***raio espetral*** *é dado por*

$$\rho(A) := \max_{i=1,\ldots,n} \{|\lambda_i| : \lambda_i \text{ é valor próprio de } A\} . \tag{4.9}$$

A norma matricial (4.6) *denomina-se por* ***norma das colunas****, enquanto que a norma matricial* (4.7) *denomina-se por* ***norma das linhas*** *.*

Note-se que se A é uma matriz hermitiana, isto é, se $A = A^*$, então

$$\|A\|_2 = \rho(A). \tag{4.10}$$

Demonstração. Comecemos pela norma $\|.\|_1$. Temos que

$$\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| \quad \text{(Desigualdade triangular)} \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}||x_j| \\
&= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \\
&\leq \left(\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| \\
&= \max_{j=1,\ldots,n} \left(\sum_{i=1}^{n} |a_{ij}| \right) \|x\|_1
\end{aligned}$$

logo pela definição (3.2) temos que

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

Por outro lado, escolhendo um vector x com todas as entradas nulas à exceção de $x_k = 1$, em que a coluna k satisfaz

$$\sum_{i=1}^{n} |a_{ik}| = \max_{j=1,\ldots,n} \left(\sum_{i=1}^{n} |a_{ij}| \right)$$

temos $\|x\|_1 = 1$ e

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| = \sum_{i=1}^{n} |a_{ik}| = \max_{j=1,\ldots,n} \left(\sum_{i=1}^{n} |a_{ij}| \right)$$

pelo que mais uma vez pela definição (3.2) temos que

$$\|Ax\|_1 \geq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|,$$

pelo que concluímos que a norma induzida pela norma vetorial (4.3) é dada por (4.6).

Para concluir (4.7) o processo é semelhante. Começamos por mostrar que

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\ &= \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}x_j| \\ &\leq \max_{j=1,\dots,n} |x_j| \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \\ &= \max_{i=1,\dots,n} \left(\sum_{j=1}^{n} |a_{ij}| \right) \|x\|_\infty\end{aligned}$$

logo

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

Escolhendo o caso particular em que o vector x tem entradas $x_j = \overline{a_{kj}}/|a_{kj}|$ (com $x_j = 0$ caso $a_{k,j} = 0$), em que k é tal que,

$$\sum_{j=1}^{n} |a_{kj}| = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$$

obtemos que $\|x\|_\infty = 1$ e

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\ &\geq \left| \sum_{j=1}^{n} a_{kj}x_j \right| \\ &= \sum_{j=1}^{n} |a_{kj}| \\ &= \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|,\end{aligned}$$

logo temos que a norma induzida pela norma vetorial (4.4) é dada por (4.7).

Falta apenas mostrar o resultado para a norma (4.7) induzida pelo produto interno euclidiano $(.,.)$. Nesse caso

$$\|Ax\|^2 = (Ax, Ax) = (x, A^*Ax), \forall x \in \mathbb{C}^n.$$

Como A^*A é uma matriz hermitiana, é definida positiva, pelo que todos os seus valores próprios λ_k são não-negativos e os seus vectores próprios v_k são uma base ortonormal do espaço [6]. Representado $x \in \mathbb{C}$ nesta base

$$x = \sum_{k=1}^{n} \alpha_k v_k$$

temos

$$\|Ax\|_2^2 = (x, A^*Ax) = \left(\sum_{j=1}^{n} \alpha_j v_j, \sum_{k=1}^{n} \alpha_k \mu_k v_k \right) = \sum_{k=1}^{n} \lambda_j |\alpha_j|^2$$

e

$$\|x\|_2^2 = (x, x) = \sum_{k=1}^{n} |\alpha_j|^2.$$

Assim é claro que

$$\|Ax\|_2^2 = \sum_{k=1}^{n} \lambda_j |\alpha_j|^2 \leq \rho\left(A^*A\right) \|x\|_2^2.$$

Por outro lado, escolhendo m tal que $|\lambda_m| = \rho\left(A^*A\right)$, temos pelos cálculos anteriores para $x = v_m$ (e portanto verificando $\|v_m\| = 1$) que

$$||Av_m||^2 = \lambda_m |\alpha_m|^2 = \lambda_m \|v_m\|_2^2 = \rho\left(A^*A\right) \|v_m\|_2^2.$$

Finalmente, falta apenas mostrar que para uma matriz hermitiana temos

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = \sqrt{\rho(A)^2} = \rho(A).$$

□

Exercício 4.3. Calcule a norma das linhas e das colunas das seguintes matrizes:

(a) $A = \begin{bmatrix} 1 & -1 \\ -2 & 3 \end{bmatrix}$;

(b) $B = \begin{bmatrix} 5 & 0 & -2 \\ 3 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}$;

(c) $C = \begin{bmatrix} i & 1 & 2 \\ 0 & 3-4i & 1 \\ 1 & 3i & -6+8i \end{bmatrix}$;

Resposta.

(a) $\|A\|_1 = 4; \quad \|A\|_\infty = 5;$

(b) $\|B\|_1 = 9; \quad \|B\|_\infty = 7;$

(c) $\|C\|_1 = 13; \quad \|C\|_\infty = 14;$

Exercício Octave 4.4. Verifique o resultado da exercício anterior utilizando o comando `norm(A,p)` em Octave para calcular a norma p (p=1,2,Inf) da matriz A.

Exercício 4.5. Mostre que a norma de matrizes $\|.\|_2$ induzida pela norma vetorial euclidiana satisfaz

$$\|A\|_2 \leq \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

Resolução.

Temos, pela desigualdade de Cauchy-Schwartz (3.7)

$$\begin{aligned} \|Ax\|_2^2 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|^2 \\ &\leq \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}| \right)^2 \left(\sum_{j=1}^{n} |x_j| \right)^2 \quad \text{(des. Cauchy-Schwartz (3.7))} \\ &= \left(\sum_{j=1}^{n} |x_j| \right)^2 \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}| \right)^2 \\ &= \left[\sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}| \right)^2 \right] \|x\|_2^2 \end{aligned}$$

logo temos o resultado pela definição de norma de um operador (3.2)

Teorema 4.6.

Para qualquer norma em $\mathbb{C}^n$ e qualquer matriz quadrada A, temos

$$\rho(A) \leq \|A\|. \tag{4.11}$$

Por outro lado, para qualquer $\varepsilon > 0$ existe uma norma tal que

$$\|A\| \leq \rho(A) + \varepsilon. \tag{4.12}$$

Demonstração. Seja v_m um vector próprio de A associado ao vector próprio tal que $|\lambda_m| = \rho(A)$. Sem perda de generalidade, assumimos v_m tem norma unitária. Então

$$||A|| = \sup_{||x||=1} \|Ax\| \geq \|Av_m\| = |\lambda_m| = \rho(A).$$

logo (4.11) está provado. Remetemos a prova de (4.12) para por exemplo [5, 10]. □

4.2 Condicionamento de sistemas lineares

Ainda que o nosso interesse principal no âmbito deste texto seja o estudo de sistemas lineares da forma (4.1), em que o operador linear A tem dimensão finita (ou seja, pode ser representado por uma matriz), o que se segue também é aplicável para operadores lineares de dimensão infinita. Assim, formulamos os resultados desta secção para operadores lineares A segundo a definição 3.4 em espaços normados, uma vez que precisamos da definição de uma norma para A. Obviamente que os resultados são também válidos no caso de A ser uma matriz.

O seguinte teorema dá-nos uma estimativa para o condicionamento de equações lineares. Note-se que para se falar em condicionamento a equação linear (4.1) deverá ter uma única solução, ou seja, o operador A deverá ser invertível no sentido da definição 3.10

Definição 4.7 (Número de condição).

Seja $A : X \to Y$ um operador invertível. Chama-se **número de condição** do operador A a

$$\text{cond}(A) = \|A^{-1}\| \, \|A\|. \tag{4.13}$$

Note-se que o número de condição depende da norma considerada.

Exercício Octave 4.8. Calcule o número de condição para as matrizes do exercício 4.3 considerando as normas das linhas, das colunas e

- usando o comando `inverse(A)` em Octave para calcular a inversa da matriz A e o comando `norm(A,p)` para calcular a norma p (p=1,2,Inf) da matriz A.
- usando o comando `cond(A,p)` em Octave para calcular o número de condição de A correspondente à norma p (p=1,2,Inf).

Resposta.

(a) $\mathrm{cond}_1(A) = 20; \quad \mathrm{cond}_2(A) \approx 14.993; \quad \mathrm{cond}_\infty(A) = 20.$

(b) $\mathrm{cond}_1(B) = 81; \quad \mathrm{cond}_2(B) \approx 37.734; \quad \mathrm{cond}_\infty(B) = 49.$

(c) $\mathrm{cond}_1(C) \approx 12.643; \quad \mathrm{cond}_2(C) \approx 9.8231; \quad \mathrm{cond}_\infty(C) \approx 16.117.$

Note-se que, independentemente da norma a que está associado o número de condição, temos o seguinte resultado.

Teorema 4.9.
Seja $A : X \to Y$ um operador invertível. Então

$$\mathrm{cond}(A) \geq 1. \tag{4.14}$$

Demonstração. Temos

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

□

Tendo em mente a teoria do erro no capítulo anterior, é importante estabelecer estimativas de erro para a resolução se sistemas lineares da forma

$$Ax = b,$$

quando o operador linear A ou o segundo membro b está afetado por erros. temos então o seguinte teorema.

Teorema 4.10 (Condicionamento de equações lineares).
Sejam X e Y espaços de Banach e seja $A : X \to Y$ um operador linear com inverso limitado $A^{-1} : Y \to X$. Seja ainda $\tilde{A}$ uma aproximação do operador A tal que

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

e sejam x e $\tilde{x}$, respetivamente, as soluções das equações lineares

$$Ax = b \qquad e \qquad \tilde{A}\tilde{x} = \tilde{b}.$$

Então temos a estimativa para o erro relativo da solução

$$\delta_{\tilde{x}} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\delta_{\tilde{A}}} \left(\delta_{\tilde{b}} + \delta_{\tilde{A}}\right) \tag{4.15}$$

em que o erro relativo é definido por (2.2).

Demonstração. Começamos por notar que pelo teorema 3.11, o operador $I + A^{-1}(\tilde{A} - A)$ é invertível. Assim concluímos que o operador $\tilde{A}$ também é invertível, uma vez que pode ser escrito como

$$\tilde{A} = A\left(I + A^{-1}(\tilde{A} - A)\right)$$

e que a sua inversa dada por

$$\tilde{A}^{-1} = \left(I + A^{-1}(\tilde{A} - A)\right)^{-1} A^{-1}$$

é majorada por

$$\|\tilde{A}^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\tilde{A} - A\|}. \tag{4.16}$$

Pela definição de x e $\tilde{x}$ temos

$$\tilde{A}(\tilde{x} - x) = \tilde{b} - b - (\tilde{A} - A)x,$$

logo aplicando a inversa de $\tilde{A}$ e majorando as normas obtemos

$$\|\tilde{x} - x\| \leq \|\tilde{A}^{-1}\|\left(\|\tilde{b} - b\| + \|\tilde{A} - A\|\|x\|\right).$$

Aplicando (4.16), dividindo ambos os membros por $\|x\|$ e considerando por definição de número de condição que $\|\tilde{A}^{-1}\| = \mathrm{cond}(A)/\|\tilde{A}\|$, temos

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\dfrac{\|\tilde{A} - A\|}{\|A\|}}\left(\frac{\|\tilde{b} - b\|}{\|A\|\|x\|} + \frac{\|\tilde{A} - A\|}{\|A\|}\right).$$

logo obtemos o resultado, uma vez que

$$\|A\|\|x\| \geq \|Ax\| = \|b\|.$$

□

O teorema anterior mostra que quanto maior for o número de condição de A, pior condicionada é a resolução da equação linear $Ax = b$, uma vez que o termo

$$\frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\delta_{\tilde{A}}} \to \infty$$

quando $\mathrm{cond}(A) \to \infty$. Este agravamento é ainda mais claro caso não existam erros na matriz A e apenas no segundo membro b.

Exercício 4.11. Mostre que caso não existam erros na matriz A invertível (i.e., $\tilde{A} = A$), temos a estimativa

$$\delta_{\tilde{x}} \leq \text{cond}(A)\delta_{\tilde{b}}. \tag{4.17}$$

Resolução.

Sai diretamente da estimativa (4.15) considerando $\tilde{A} - A = 0$.

Exercício Octave 4.12. Considere a norma $\|.\|_2$ e o sistema linear $Ax = b$ com

$$A = \begin{bmatrix} 3 & 5 & -1 \\ 2 & -4 & 2 \\ 5 & 1 & 1.0001 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \\ 10 \end{bmatrix}.$$

(a) Calcule o número de condição de A, usando o comando `cond` do Octave.

(b) Que conclui sobre o condicionamento do sistema $Ax = b$.

(c) Considere agora as aproximações

$$\tilde{A} = \begin{bmatrix} 3 & 5 & -1 \\ 2 & -4 & 2 \\ 5 & 1 & 1.00005 \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} 0 \\ 1 \\ 10.00005 \end{bmatrix}.$$

Calcule os erros relativos da matriz $\tilde{A}$ e do vector $\tilde{b}$, pela sua definição utilizando o comando `norm`.

(d) Verifique se está nas condições do teorema 4.10, utilizando os comandos `inv` e `norm` do Octave.

(e) Calcule a estimativa para o erro relativo da solução do sistema $\tilde{A}\tilde{x} = \tilde{b}$ em relação à solução do sistema inicial $Ax = b$.

(f) Calcule o erro relativo efetivamente cometido.

Resposta.

(a) $\text{cond}_2(A) = 1.3831 \times 10^5$.

(b) Como o número de condição é elevado, o sistema é mal condicionado.

(c) $\delta_{\tilde{A}} = \dfrac{\|\tilde{A} - A\|_2}{\|A\|_2} \approx 6.8778 \times 10^{-6}; \delta_{\tilde{b}} = \dfrac{\|\tilde{b} - b\|_2}{\|b\|_2} = 4.9752 \times 10^{-6}$.

(d) Basta verificar que

$$\|A^{-1}\|_2\|\tilde{A} - A\|_2 = 0.95130 < 1,$$

logo estamos nas condições do teorema.

(e) Por (4.15) temos $\delta_{\tilde{x}} \leq 33.661$, ou seja, a estimativa de erro diz-nos que o erro cometido é no máximo $3366,1\%$.

(f) Temos as soluções

$$x = [-2.4545, 3.2727, 9]^T \times 10^4; \quad \tilde{x} = [-4.9091, 6.5455, 18]^T \times 10^4$$

logo obtemos $\delta_{\tilde{x}} = 1$, ou seja, um erro de 100 %.

Vamo-nos agora concentrar na resolução de sistemas da forma

$$Ax = \tilde{b}$$

em que o valor $\tilde{b}$ é uma pequena perturbação do valor exato b, isto é, o segundo membro da equação está afetado de erros. Como vimos, o condicionamento do sistema está relacionado com o número de condição de A, pela equação (4.17).

Note-se também que o número de condição está também relacionado com a invertibilidade de A. Em particular, se A tiver valores próprios próximos de zero, a invertibilidade de A fica mal condicionada e logo o seu número de condição será grande.Esta afirmação tem como suporte a definição de número de condição e o teorema 4.6. De facto, como os valores próprios de A^{-1} são os inversos de A, temos que se A tem um valor próprio muito próximo de zero em relação aos restantes, então A^{-1} tem um valor próprio muito grande, logo $\rho(A^{-1})$ é muito grande. Desta forma, pelo teorema 4.6, qualquer que seja a norma considerada, temos $\|A^{-1}\|$ muito grande e logo o número de condição será muito grande (partindo do princípio que $\|A\|$ não é próximo de zero, o que implicaria que os seus valores próprios fossem todos próximos de zero, mais uma vez pelo teorema 4.6).

Assim, uma forma de controlar o condicionamento da resolução de uma equação linear $Ax = b$ é controlar os valores próprios de A, ou seja, resolver uma equação linear semelhante em que os valores próprios do operador linear considerado sejam todos afastados de zero.

4.3 Regularização

Nesta secção vamos ilustrar alguns métodos de regularização de sistemas mal condicionados, ou seja, vamos considerar processos para estabilizar resoluções de equações lineares. Começamos por recordar o seguinte resultado de Álgebra Linear ou Análise Funcional que nos será útil para o nosso objetivo.

Teorema 4.13 (Sistema Singular).

Seja A uma matriz $m \times n$ com característica r. Então existem sistemas ortonormais $u_1, u_2, \ldots, u_n \in \mathbb{C}^n$ e $v_1, v_2, \ldots, v_m \in \mathbb{C}^m$ e valores reais

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > \mu_{r+1} = \ldots \mu_n = 0$$

tais que

$$Au_j = \mu_j v_j, \qquad A^* v_j = \mu_j u_j, \quad j = 1, \ldots, r \tag{4.18}$$

$$Au_j = 0, \qquad j = r+1, \ldots, n \tag{4.19}$$

$$A^* v_j = 0, \qquad j = r+1, \ldots, m. \tag{4.20}$$

*Ao sistema (μ_j, u_j, v_j) chama-se **sistema singular** de A e aos valores μ_j chamam-se **valores singulares** de A .*

Mais ainda, o sistema linear $Ax = b$ tem solução única se e só se

$$(b, z) = 0, \quad \forall z \in \mathbb{C}^m : A^* z = 0.$$

Nesse caso, a solução é dada por

$$x = \sum_{j=1}^{r} \frac{1}{\mu_j} (b, v_j) u_j. \tag{4.21}$$

Demonstração. O resultado sai fora do âmbito deste texto, remetemos para [5]. □

Exercício Octave 4.14. Utilize o comando `[V MU U] = svd(A)` para obter o sistema singular da matriz $\tilde{A}$ dada no exercício 4.12. Verifique que as condições (4.18)-(4.20) são satisfeitas.

Resposta.

Obtemos os vectores coluna da matriz `U`

$$u_1 = [-0.59153, -0.79603, 0.12814]^T; \quad u_2 = [-0.76710, 0.50670, -0.39347]^T;$$
$$u_3 = [-0.24829, 0.33104, 0.91036]^T;$$

e os vectores coluna de `V`

$$v_1 = [-0.80923, 0.31051 - 0.49872]^T; \quad v_2 = [0.10866, -0.75515, -0.64649]^T;$$
$$v_3 = [-0.57735, -0.57735, 0.57735]^T;$$

que prefazem um sistema ortonormal e obtemos os valores singulares a partir da diagonal da matriz `MU`

$$\mu_1 = 7.2697; \quad \mu_2 = 5.7577; \quad \mu_3 = 2.6280 \times 10^{-5}.$$

A expressão da solução (4.21) ilustra a relação entre os valores singulares de A e o condicionamento da equação $Ax = b$. Para simplificar a análise seguinte, vamos considerar que a matriz A é quadrada de dimensões $n \times n$ e invertível (de forma a que o sistema linear tenha solução), o que faz com que todos os seus valores próprios e valores singulares sejam distintos de zero.

Começamos por notar que se os valores singulares μ_j forem próximos de zero, o fator $1/\mu_j$ torna a solução numérica instável. Dito de outra forma, como μ_j^2 são os valores próprios de A^*A pois de (4.18) temos

$$A^*Au_j = \mu_j A^* v_j = \mu_j^2 u_j,$$

temos por (4.10) que

$$\mathrm{cond}_2(A) = \frac{\mu_1^2}{\mu_n^2}$$

logo se o menor valor singular μ_n for muito próximo de zero, o bom condicionamento do sistema $Ax = b$ fica comprometido.

Finalmente, falta apenas comparar os valores próprios λ_j de A com os seus valores singulares μ_j. Começamos por notar que os valores próprios de matriz transposta conjugada A^* são os valores próprios conjugados de A. Assim, sabendo que o determinante de uma matriz é dada pelo produto dos seus valores próprios, temos

$$\prod_{j=1}^{n} \mu_j^2 = \det(A^*A) = \det(A^*)\det(A) = \prod_{j=1}^{n} \bar{\lambda}_j^2 \prod_{j=1}^{n} \lambda_j^2 = \prod_{j=1}^{n} |\lambda_j|^2.$$

Concluímos então que se um dos valores próprios for próximo de zero (tornando a inversão da matriz pouco estável), o produto dos valores próprios e logo dos valores singulares também será próximo de zero, pelo que pelo menos um dos valores singulares será próximo de zero. Tal como tínhamos visto anteriormente, se os valores próprios de A forem próximos de zero, o condicionamento da equação linear associada é mau.

4.3.1 Regularização por decomposição em valores singulares

A equação (4.21) ilustra o mau condicionamento gerado por valores singulares μ_j próximos de zero. Por outro lado, também permite obter uma forma de aproximar a solução do sistema linear de forma estável. Para isso basta truncar o somatório, não considerando valores singulares μ_j menores que determinado $\varepsilon > 0$ arbitrário. Desta forma a contribuição do fator $1/\mu_j$ desaparece para os valores instáveis. Ficamos então com a aproximação da solução dada por

$$x_p = \sum_{j=1}^{p} \frac{1}{\mu_j}(b, v_j)u_j,$$

em que p é escolhido como o índice tal que $\mu_p \geq \varepsilon$ e $\mu_{p+1} < \varepsilon$. A escolha do valor de ε ou de outra forma, do índice p é crucial para a qualidade da aproximação. Por um lado, queremos uma solução estável, mas por outro não queremos truncar demasiado a série de forma a que o erro cometido não seja muito grande, uma vez que pela desigualdade triangular temos

$$\|x - x_p\|_2 \leq \sum_{j=p+1}^{n} \frac{1}{\mu_j}|(b, v_j)|.$$

A fórmula do erro mostra-nos que caso b seja perpendicular ao espaço gerado pelos vectores v_j para $j = p + 1, \ldots, n$, temos $(b, v_j) = 0$ e logo o erro cometido é zero. Por outro lado, se b pertencer ao espaço gerado por esses vectores, o erro pode ser muito grande.

4.3.2 Regularização de Tikhonov

A regularização por decomposição em valores singulares é pouco prática, porque pressupõe calcular-se o sistema singular de A o que muitas vezes só por si é uma tarefa morosa e até pouco estável. Nesta secção ilustramos outra forma de fazer regularização, nomeadamente a regularização de Tikhonov. A ideia é substituir a resolução do sistema

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

pela do sistema

$$(\alpha I + A^*A)x_\alpha = A^*\tilde{b} \tag{4.22}$$

para uma constante de regularização $\alpha > 0$. Como iremos mostrar de seguida, o sistema 4.22 é melhor condicionado que o original e a sua solução x_α é próxima de x.

Teorema 4.15 (Regularização de Tikhonov).

Seja (μ_j, u_j, v_j) o sistema singular de uma matriz A e seja $\alpha > 0$ uma constante real.

Então a matriz $M = \alpha I + A^*A$ relativa ao sistema (4.22) *de regularização de Tikhonov tem número de condição*

$$\operatorname{cond}_2(M) \leq \frac{\mu_1^2 + \alpha}{\mu_n^2 + \alpha}. \tag{4.23}$$

Além disso a solução de (4.22) *pode ser escrita na forma*

$$x_\alpha = \sum_{j=1}^{n} \frac{\mu_j}{\alpha + \mu_j^2} (\tilde{b}, v_j) u_j. \tag{4.24}$$

Se x é a solução do sistema $Ax = b$ então temos a estimativa de erro

$$\|x - x_\alpha\|_2 \leq \|b\|_2 \sum_{j=1}^{n} \left| \frac{\alpha}{\mu_j(\alpha + \mu_j^2)} \right| + e_{\tilde{b}} \sum_{j=1}^{n} \left| \frac{\mu_j}{(\alpha + \mu_j^2)} \right| \tag{4.25}$$

em que o erro absoluto do segundo membro é dado por $e_{\tilde{b}} = \|b - \tilde{b}\|_2$.

Demonstração. Por (4.18) temos que

$$(\alpha I + A^*A) u_j = (\alpha + \mu_j^2) u_j,$$

logo os valores próprios de M são dados por $(\alpha + \mu_j^2)$. Mais ainda, como M é hermitiana, o resultado (4.23) sai diretamente de (4.10). A representação (4.24) sai diretamente da representação singular (4.21) considerando que $(\alpha + \mu_j^2, u_j, u_j)$ é um sistema singular de M, considerando que $(A^*\tilde{b}, u_j) = (\tilde{b}, Au_j) = \mu_j(\tilde{b}, v_j)$. Finalmente, considerando as representações (4.21) e (4.24) temos

$$\begin{aligned}
x - x_\alpha &= \sum_{j=1}^{n} \left(\frac{1}{\mu_j}(b, v_j) - \frac{\mu_j}{(\alpha + \mu_j^2)}(\tilde{b}, v_j) \right) u_j \\
&= \sum_{j=1}^{n} \left(\left[\frac{1}{\mu_j} - \frac{\mu_j}{(\alpha + \mu_j^2)} \right] (b, v_j) + \frac{\mu_j}{(\alpha + \mu_j^2)}(b - \tilde{b}, v_j) \right) u_j \\
&= \sum_{j=1}^{n} \left(\frac{\alpha}{\mu_j(\alpha + \mu_j^2)}(b, v_j) + \frac{\mu_j}{(\alpha + \mu_j^2)}(b - \tilde{b}, v_j) \right) u_j
\end{aligned}$$

e pela aplicação da desigualdade triangular e de Cauchy, obtemos a estimativa (4.25).

□

A aplicação da regularização de Tikhonov tem uma grande vantagem em relação à regularização por decomposição em valores singulares: não é necessário o cálculo do sistema singular. De facto, para aplicar a regularização de Tikhonov basta resolver o sistema (4.22). A grande desvantagem da regularização de Tikhonov é que é necessário uma escolha apropriada da constante de regularização α, o que nem sempre é trivial. Por um lado α deve ser pequeno, para garantir que o sistema linear a resolver não é muito alterado. Este facto é claro, uma vez que a primeira parcela do segundo membro da estimativa de erro (4.25) é tanto mais pequena, quanto mais pequeno for α. Por outro lado, α não pode ser demasiado pequeno para garantir que erro de $\tilde{b}$ não é amplificado, conforme se verifica na segunda parcela da estimativa (4.25). Um indicador semelhante é também dado pelo majorante para o número de condição (4.23), que da mesma forma indica que α não pode ser muito pequeno, para garantir que um bom condicionamento da matriz do sistema de Tikhonov (4.22).

Exercício Octave 4.16. Considere a matriz A de dimensões $n \times n$, cuja entrada (j, k) é dada por $a_{jk} = \sin(jk/n)$ e o vector coluna b tal que $b_j = \sum_{k=1}^{n} k \sin(kj/n)$. Considere ainda o vector $\tilde{b}$ afetado de erro, tal que

$$\tilde{b}_j = b_j + 10^{-12} \cos(2000j/n).$$

Sabe-se que a solução do sistema $Ax = b$ é dada pelo vector com componente $x_j = j,\ j = 1, 2, \ldots, N$. Considere nas alíneas seguintes n=12.

(a) Construa em Octave a matriz A e os vectores b e $\tilde{b}$.

(b) Determine o erro relativo e absoluto de $\tilde{b}$ na norma euclideana, utilizando o comando `norm` em Octave.

(c) Determine o número de condição de A na norma euclideana, utilizando o comando `cond`.

(d) Determine o erro relativo da solução do sistema $A\tilde{x} = \tilde{b}$ obtida pela resolução direta xtil = A\ btil, em relação à solução x do sistema sem erros.

(e) Determine a aproximação por decomposição em valores singulares, utilizando os $p = 10$ maiores valores singulares. Calcule o erro relativo cometido.

(f) Determine a aproximação por regularização de Tikhonov, utilizando a constante de regularização $\alpha = 10^{-10}$. Calcule o erro relativo cometido.

(g) Compare graficamente as aproximações obtidas com a solução do sistema, comentando o resultado obtido.

Resposta.

(a) Num ficheiro '.m', escrevemos o seguinte algoritmo:

```
N=12;
b = double(zeros(N,1));
btil = double(zeros(N,1));
A = double(zeros(N,N));
for j= 1:  N
  for k= 1:  N
    A(j,k) = sin(j*k/N); %construção da matriz A
    b(j) = b(j) + k*sin(k*j/N); %construção de b
  end
  btil(j)= b(j)+10^(-12)*cos(2000*j/N); %const.  de btil
end
```

(b) Temos os erros absoluto $e_{\tilde{b}} = 2.1277 \times 10^{-12}$ e relativo $\delta_{\tilde{b}} = 1.9864 \times 10^{-14}\%$ nos dados, sendo portanto erros muito pequenos.

(c) Temos $\mathrm{cond}_2(A) = 1.7470 \times 10^{14}$, logo o sistema $Ax = b$ é muito mal condicionado.

(d) Temos $\tilde{x} = [27.18, -40.25, 47.18, -31.27 \ldots, 12.00]$ logo obtemos

$$\delta_{\tilde{x}} = 312.10\%.$$

Este erro enorme nos resultados a partir de um erro minúsculo nos dados é justificado pelo mau condicionamento do sistema e pela não regularização da sua solução.

(e) Obtemos a solução aproximada

$$x_p \approx [1.00, 2.00, 3.00, \ldots, 12.00]^T$$

escrevendo num ficheiro '.m' o seguinte algoritmo:

```
 [v mu u] = svd(A);
p = 10;
xp = zeros(N,1);
for j= 1:  p
```

```
  xp = xp + dot(btil,v(:,j))*u(:,j)/mu(j,j);
end
```

O erro relativo é $\delta_{x_p} = 0.00016850\%$.

(f) Obtemos a solução

$$x_\alpha \approx [1.00, 2.00, 3.00, \ldots, 12.00]^T$$

escrevendo num ficheiro '.m' o seguinte algoritmo:

```
alpha = 10^(-10);
MTik = alpha*eye(N) + A'*A;
bTik = (A')*btil;
xTik = MTik\ bTik;
```

O erro relativo é $\delta_{x_\alpha} = 0.00069637\%$.

(g) A comparação das três soluções aproximadas com a solução exata é ilustrada graficamente na figura 4.1, com a sequência de comandos:

```
figure(1)
plot(x,'r-');
hold on;
plot(xtil,'b-');
plot(xp,'g-');
plot(xTik,'c-');
legend('Solucao exata','Solucao com erro','SVD','Tikhonov');
hold off;
```

A instabilidade da solução do sistema não regularizado é óbvia, enquanto que as soluções regularizadas coincidem graficamente com a solução exata neste caso.

O condicionamento de uma equação linear e a sua regularização dependem portanto do sistema singular da matriz A associada. Desta forma, é necessário ter resultados para que rapidamente se consigam localizar os valores próprios de uma matriz, o que para matrizes de grandes dimensões pode ser um processo moroso. A próxima secção dedica-se ao estudo deste problema.

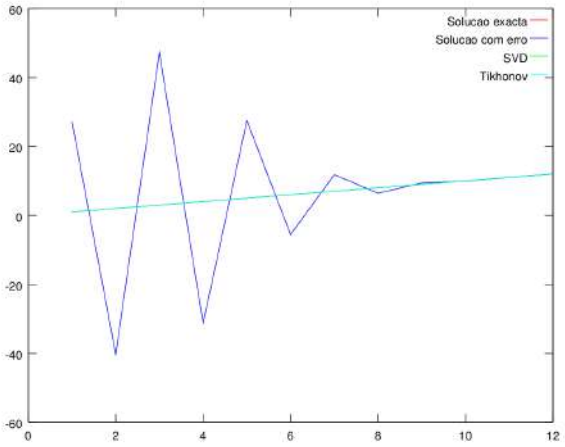


Figura 4.1: *Comparação das soluções aproximadas e exata do exercício 4.16.*

4.4 Localização de Valores próprios

Dedicamos esta secção ao estudo de métodos que permitam localizar valores próprios de operadores lineares. Como sabemos por definição, diz-se que λ é um valor próprio do operador linear A se existe um $v \neq 0$ tal que

$$Av = \lambda v.$$

No caso de A ser uma matriz quadrada $n \times n$, sabemos que os valores próprios são as raízes do polinómio característico de grau n definido por [6].

$$p_A(\lambda) = \det(A - \lambda I).$$

Consegue-se mostrar que as raízes do polinómio dependem continuamente dos seus coeficientes, pelo que o problema de determinar os valores próprios de uma matriz de forma algébrica depende continuamente dos dados. No entanto, para matrizes de grandes dimensões, este processo é demasiado moroso, pelo que houve a necessidade de considerar formas mais expeditas de o fazer.

Uma delas é considerar o Teorema de Gershgorin, que embora não indique valores aproximados para os valores próprios, determina enquadramentos no plano complexo onde os valores próprios se podem localizar.

4.4.1 Teorema de Gershgorin

Passamos a enunciar o teorema.

Teorema 4.17 (Teorema de Gershgorin).

Seja A uma matriz A quadrada de dimensões $n \times n$ com entradas $a_{ij},\ i, j, = 1, 2, \ldots, n$. Seja λ um valor próprio de A.

Então

$$\lambda \in \left(\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, r_i^{(\ell)}\right) \right) \quad e \quad \lambda \in \left(\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, r_i^{(c)}\right) \right), \tag{4.26}$$

em que $\bar{B}(a_{ii}, r_i)$ representa a bola fechada de centro no elemento da diagonal a_{ii} e raio r_i, com

$$r_i^{(\ell)} = \sum_{j \neq i} |a_{ij}| \text{ (linhas)} \quad ou \quad r_i^{(c)} = \sum_{j \neq i} |a_{ji}| \text{ (colunas)}$$

Além disso, se a reunião $\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, r_i^{(\ell)}\right)$ das bolas pela análise das linhas tem $k \leq n$ componentes conexas, cada uma formada pela reunião de $m_1, m_2, \ldots m_k$ bolas então em cada uma dessas componentes conexas existem exatamente $m_1, m_2, \ldots m_k$ valores próprios de A, respetivamente.

O mesmo é válido para as colunas.

Demonstração. Fazemos apenas a prova no caso das linhas, reparando que o caso das colunas fica automaticamente provado uma vez que os valores próprios de A^T coincidem com os de A.

Seja v um vector próprio de A e seja o índice k tal que $|v_k| = \max_{i=1,\ldots,n} |v_i|$. Então, para a componente k temos que

$$\lambda v_k = [Av]_k = \sum_{j=1}^{n} a_{kj} v_j$$

logo

$$(\lambda - a_{kk}) v_k = \sum_{j \neq k} a_{kj} v_j.$$

Assim, dividindo por v_k (que é diferente de zero por definição de vector próprio)

$$(\lambda - a_{kk}) = \sum_{j \neq k} a_{kj} \frac{v_j}{v_k}.$$

e aplicando a desigualdade triangular, temos

$$|\lambda - a_{kk}| = \sum_{j \neq k} |a_{kj}| \underbrace{\left| \frac{v_j}{v_k} \right|}_{\leq 1}.$$

Assim existe um k tal que $\lambda \in \bar{B}\left(a_{kk}, r_k^{(\ell)}\right)$, como queríamos demonstrar.

Para demonstrar a segunda parte do teorema, começamos por considerar

$$A_t = D + t(A - D), \quad t \in [0, 1]$$

em que D é a matriz diagonal que coincide com a diagonal de A. Assim, a diagonal de A_t coincide com a diagonal de A para qualquer $t \in [0, 1]$. Mais ainda, por aplicação de (4.26) sabemos que o valor próprio $\lambda_j^{(t)}$ de A_t satisfaz

$$\lambda_j^{(t)} \in \left(\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, t\, r_i^{(\ell)}\right)\right) \subset \left(\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, r_i^{(\ell)}\right)\right)$$

e como A_t é contínua em t, os seus valores próprios dependem continuamente de t, pelo que o conjunto

$$\Lambda_j = \{\lambda_j^{(t)} : t \in [0, 1]\}$$

é conexo, contém a_{jj} (pois $\lambda_j^{(0)} = a_{jj}$) e logo tem de pertencer à componente conexa de $\left(\bigcup_{i=1}^{n} \bar{B}\left(a_{ii}, t\, r_i^{(\ell)}\right)\right)$ que contém a_{jj}. □

O teorema de Gershgorin permite localizar os valores próprios de A sem grande custo computacional. Na realidade, para uma matriz $n \times n$ basta fazer $2n(n-1)$ somas para obter os raios das circunferências, enquanto que um processo como a eliminação de Gauss para determinar os valores próprios tem uma complexidade da ordem de n^3 operações. O preço a pagar é que no caso do Teorema de Gershgorin não obtemos os valores exatos dos valores próprios, apenas obtemos aproximações da sua localização. No entanto, ainda assim podemos ter informação valiosa a partir do teorema de Gershgorin, conforme se ilustra no exercício seguinte.

Exercício 4.18. Seja A uma matriz $n \times n$ com diagonal $a_{ii} = 1$, $i = 1, 2, \ldots, n$ e restantes elementos $a_{ij} = 1/n$, $i \neq j$. Mostre que A é invertível.

Resolução.

Pelo Teorema de Gershgorin, quer pela análise de linhas quer pela análise de colunas, concluímos que todos os valores próprios de A estão contidos na bola $\bar{B}(1, (n-1)/n)$. Como a origem não está contida nessa bola, zero não é valor próprio de A, logo A é invertível.

Note-se que a informação que obtemos da análise por linhas é potencialmente diferente da que obtemos pela análise por colunas. A interseção da informação que se obtém por linhas e colunas leva muitas vezes a resultados mais precisos do que as análises isoladas. Alguns exemplos são demonstrados no exercício seguinte.

Exercício Octave 4.19. Faça um algoritmo em Octave que, dada uma matriz A quadrada $n \times n$ devolva as bolas que contêm os seus valores próprios, pelo teorema de Gershgorin.

Resolução.

A proposta seguinte calcula também os valores próprios de A e coloca-os no gráfico. Se A tiver grandes dimensões as linhas de código correspondentes a este cálculo devem ser eliminadas por questões de tempo computacional.

```
function [lc, lr] = Gersh(A)
% Traça os circulos de Gershgorin.
% INPUT: A- Matriz quadrada
% OUTPUT: lc - centro das bolas no plano complexo;
% lr - raio das bolas;
%   primeira coluna (analise colunas);
%   segunda coluna (analise linhas)
N =size(A,1);
if size(A,2) ~= N
  disp('A matriz nao e quadrada!')
  return
end
lc = diag(A);
lr = zeros(N,2);
lr(:,1) = sum(abs(A),1).'-abs(lc);
lr(:,2) = sum(abs(A),2)-abs(lc);
figure(1); hold on;
[ evec, eval ] = eig(A);
lt = 0:  pi/30:  2*pi;
for j = 1:N
  plot(real(lc(j))+lr(j,1)*cos(lt),...
    imag(lc(j))+lr(j,1)*sin(lt),'r-','Linewidth',1.5);
  plot(real(lc(j))+lr(j,2)*cos(lt),...
    imag(lc(j))+lr(j,2)*sin(lt),'g-');
  plot(real(eval(j,j)),imag(eval(j,j)),'b+');
end
legend('Colunas','Linhas','Valores Proprios')
hold off;
return
```

Exercício Octave 4.20. Utilize o algoritmo anterior para localizar os valores próprios das matrizes seguintes:

(a) $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & -5 & 3 \\ 2 & -1 & 8 \end{bmatrix}$

(b) $B = \begin{bmatrix} 7+6i & 2 & -1 & 0 \\ i & -3+7i & 1+i & 2i \\ 0 & -2 & 4-3i & 2+i \\ 1 & 1+i & 2 & -5-i \end{bmatrix}$

(c) $C = \begin{bmatrix} -i & 0 & -1 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 3+i & 0 & -1 \\ 0 & -i & 0 & -3 & 1 \\ 0 & 1 & 0 & -1 & 4i \end{bmatrix}$

Resposta.

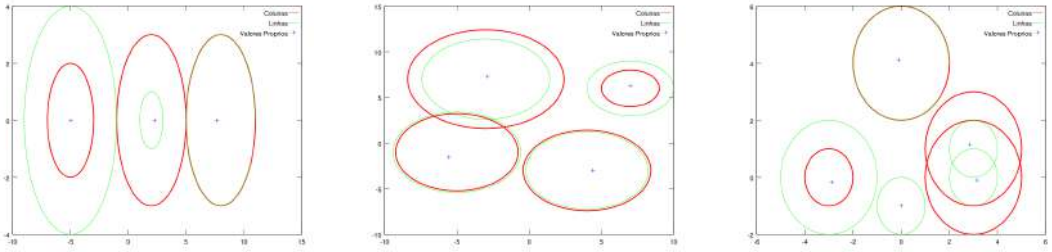


Figura 4.2: *Resposta ao exercício 4.20 alíneas (a), (b) e (c), respetivamente.*

(a) A ilustração está na figura 4.2. Pelo cruzamento da informação de linhas e colunas concluímos que um dos valores próprios está na bola $\bar{B}(-5, 2)$, outro na bola $\bar{B}(2, 1)$ e outro na bola $\bar{B}(8, 3)$. Assim, os valores próprios de A verificam $1 \leq |\lambda| \leq 11$ pelo que a matriz é invertível e o sistema bem condicionado.

(b) A ilustração está na figura 4.2. Pelo cruzamento da informação de linhas e colunas concluímos que um dos valores próprios está na bola $\bar{B}(7 + 6i, 2)$, outro na bola $\bar{B}(4 - 3i, 4.4142)$ e outros dois na componente conexa dada por

$$\left(\bar{B}(-3 + 7i, 5.42) \cup \bar{B}(-5 - i, 4.24)\right) \cap \left(\bar{B}(-3 + 7i, 4.42) \cup \bar{B}(-5 - i, 4.42)\right).$$

(c) A ilustração está na figura 4.2. Pela informação das colunas temos que um dos valores próprios é $\lambda = -i$, outro está na bola $\bar{B}(-3, 1)$ e outro na bola $\bar{B}(4i, 2)$. Pela análise de linhas temos que os restantes dois valores próprios estão na componente conexa dada pela união $\bar{B}(3, 1) \cup \bar{B}(3 + i, 1)$.

Capítulo 5

Aproximação de Funções

Passamos agora para o estudo de aproximação de funções. Pretendemos aproximar uma função real $f : \mathbb{R} \to \mathbb{R}$ de variável real, da qual conhecemos apenas alguns pontos ou propriedades. Esta é uma área da matemática aplicada com inúmeras aplicações em vários campos, uma vez que quando medimos o valor de determinada função, apenas temos acesso ao seu valor em alguns pontos e não no contínuo de um intervalo $[a, b]$.

Nesta secção estaremos interessados em três casos especiais:

(a) Conhecemos a função f n-vezes diferenciável em torno de um ponto x_0 e queremos aproximá-la num intervalo $[a, b]$ contendo x_0. O método de aproximação apropriado é a Fórmula de Taylor.

(b) Conhecemos o valor da função f em $n+1$ pontos $x_0, x_1, \ldots, x_n$ de um intervalo $[a, b]$ e queremos aproximá-la nesse intervalo. Utilizaremos interpolação polinomial.

(c) Conhecemos o valor aproximado da função f em $n+1$ pontos $x_0, x_1, \ldots, x_n$ de um intervalo $[a, b]$ assim como o seu tipo de comportamento (polinomial, sinusoidal, exponencial,...) e queremos aproximá-la nesse intervalo. Utilizaremos aproximação por mínimos quadrados.

Começamos com o primeiro caso.

5.1 Fórmula de Taylor

A fórmula de Taylor permite-nos representar uma função $f : [a, b] \to \mathbb{R}$ real de variável real e $(n + 1)$-vezes diferenciável, conhecendo apenas o valor dessa

mesma função e das suas derivadas num ponto $x_0 \in [a,b]$. Começamos por definir o polinómio de Taylor de f em torno de x_0.

Definição 5.1 (Polinómio de Taylor).

Seja $f : [a,b] \to \mathbb{R}$ uma função real de variável real $(n+1)$-vezes diferenciável.

Chama-se **polinómio de Taylor de f de ordem n em torno de x_0** ao polinómio

$$\begin{aligned} P_f^{(n)}(x) &:= \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \\ &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \end{aligned} \tag{5.1}$$

Exercício 5.2. Determine o polinómio de Taylor de ordem 3 das seguintes funções em torno dos pontos dados:

(a) $f(x) = e^x$ em torno de $x_0 = 0$;

(b) $g(x) = \cos x$ em torno de $x_0 = 0$;

(c) $h(x) = x^4$ em torno de $x_0 = 1$;

(d) $j(x) = \sqrt{x}$ em torno de $x_0 = 1$;

Resposta.

(a) $P_f^{(3)}(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$;

(b) $P_g^{(3)}(x) = 1 - \frac{x^2}{2}$;

(c) $P_h^{(3)}(x) = 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3$;

(d) $P_j^{(3)}(x) = 1 + \frac{x-1}{2} - \frac{(x-1)^2}{8} + \frac{(x-1)^3}{16}$;

O resultado principal é o que se segue.

Teorema 5.3 (Fórmula de Taylor).

Seja $f : [a, b] \to \mathbb{R}$ uma função real de variável real $(n + 1)$-vezes diferenciável e seja $x_0 \in]a, b[$.

Então, a função f pode ser representada por

$$f(x) = P_f^{(n)}(x) + R_f^{(n)}(x) \tag{5.2}$$

em que $P_f^{(n)}$ é o polinómio de Taylor (5.1) em torno de x_0 e $R_f^{(n)}$ é o resto da Fórmula de Taylor dado por

$$R_f^{(n)}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

para algum $\xi \in [a, b]$.

Demonstração. A demonstração pode ser encontrada em qualquer livro de Análise Matemática, por exemplo [7]. □

Nota 5.4. O resto $R_f^{(n)}$ da fórmula de Taylor equivale ao **erro de truncatura** da aproximação da função f pelo polinómio de Taylor $P_f^{(n)}$, uma vez que corresponde exatamente a truncar a série de Taylor de f dada por

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j$$

a partir da parcela n.

O teorema anterior diz-nos que se o resto $R_f^{(n)}(x) \approx 0$ então podemos aproximar a função f pelo seu polinómio de Taylor de ordem n, isto é,

$$f(x) \approx P_f^{(n)}(x). \tag{5.3}$$

Nota 5.5 (Série de Taylor). Caso a função f seja infinitamente diferenciável no intervalo $[a, b]$, então podemos identificar f com o seu polinómio de Taylor de ordem ∞. Neste caso, temos

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j,$$

sempre que o somatório do lado direito fizer sentido e $R_f^{(n)}(x) \to 0$ quando $n \to \infty$. Ao somatório até infinito do lado direito chama-se Série de Taylor. Desta forma, a aproximação (5.3) de f pelo seu polinómio de Taylor de ordem n acaba por ser uma truncatura da série anterior a partir da parcela de ordem $n + 1$.

Para que a aproximação (5.3) seja válida, basta garantir que o resto do polinómio de Taylor é próximo de zero. Caso as derivadas de ordem superior a n_0 sejam limitadas no intervalo $[a, b]$, ou seja,

$$|f^{(n+1)}(\xi)| < C, \quad \forall \xi \in [a, b], \forall n \geq n_0$$

então temos o seguinte:

- Se x for próximo de x_0, então $(x - x_0)^{n+1} \approx 0$ logo o resto $R_f^{(n)}(x) \approx 0$;
- Temos também que para um $x \in [a, b]$ fixo

$$\lim_{n\to\infty} |R_f^{(n)}(x)| \leq C \lim_{n\to\infty} \left| \frac{(x - x_0)^{n+1}}{(n+1)!} \right| = 0,$$

  pelo que se n for suficientemente grande então $R_f^{(n)}(x) \approx 0$.

Note-se que se as derivadas de ordem superior não forem limitadas (como é em geral o caso), nada se pode garantir. No entanto, para $|x - x_0| < 1$, a convergência do polinómio de Taylor de ordem n para f quando $n \to \infty$ será assegurada se as derivadas crescerem mais lentamente que o fatorial $n!$.

Exercício 5.6. Seja $f(x) = \sin x$.

(a) Determine o polinómio de Taylor de ordem 3 de f em torno de $x_0 = 0$.

(b) Determine a aproximação de terceira ordem de $f(1)$, recorrendo ao polinómio anterior.

(c) Determine uma estimativa de erro para a aproximação anterior, recorrendo à formula de Taylor.

Resposta.

(a) $P_f^{(3)}(x) = x - \dfrac{x^3}{6}$;

(b) $f(1) \approx P_f^{(3)}(1) = \dfrac{5}{6} \approx 0.8333$;

(c) $|f(1) - P_f^{(3)}(1)| = |R_f^{(3)}(1)| = \dfrac{|\sin \xi|}{4!}(1 - 0)^4 \overset{\xi\in[0,1]}{\leq} \dfrac{\sin(1)}{24} \approx 0.03506.$

Nota 5.7. Esta é de facto a forma como algumas calculadoras calculam o seno de um valor. Claro que o polinómio de Taylor usado é de ordem bastante superior!

Exercício 5.8. Seja $f(x) = e^{-x}$.

(a) Determine o polinómio de Taylor de ordem 2 de f em torno do ponto $x_0 = 0$.

(b) Determine a aproximação de segunda ordem de $f(0.5)$, recorrendo ao polinómio anterior.

(c) Determine uma estimativa de erro para a aproximação anterior, recorrendo à fórmula de Taylor.

Resposta.

(a) $P_f^{(2)}(x) = 1 - x + \dfrac{x^2}{2}$;

(b) $f(0.5) \approx P_f^{(2)}(0.5) = 0.625$;

(c) $|f(0.5) - P_f^{(2)}(0.5)| = |R_f^{(2)}(0.5)| = \dfrac{|-e^{-\xi}|}{3!}(0.5-0)^3 \overset{\xi\in[0,0.5]}{\leq} \dfrac{0.5^3}{6} \approx 0.02083.$

Exercício 5.9. Seja $f(x) = |x|$.

(a) Determine o polinómio de Taylor de ordem 3 de f em torno do ponto $x_0 = 1$.

(b) Calcule $P_f^{(3)}(-1)$.

(c) Indique, justificando, se este pode constituir uma boa aproximação de $f(-1)$.

Resposta.

(a) $P_f^{(3)}(x) = x$;

(b) $P_f^{(3)}(-1) = -1$;

(c) Nada se pode concluir, pois f não é diferenciável em $[-1, 1]$, logo não se pode aplicar a fórmula de Taylor.

Exercício 5.10. Seja $f(x) = \ln(x + 1)$.

(a) Determine o polinómio de Taylor de ordem 2 de f em torno do ponto $x_0 = 0$.

(b) Calcule $P_f^{(2)}(-1)$.

(c) Indique, justificando, se este pode constituir uma boa aproximação de $f(-1)$.

Resposta.

(a) $P_f^{(2)}(x) = x - \dfrac{x^2}{2}$;

(b) $P_f^{(2)}(-1) = -\dfrac{3}{2}$;

(c) Nada se pode concluir, pois f não está definida em $x = -1$, sendo que

$$\lim_{x \to -1^+} f(x) = -\infty.$$

Vamos agora introduzir uma notação que nos será útil no futuro.

Definição 5.11.

Diz-se que a função real $g = g(h)$ de variável real h é $\boldsymbol{O(h^n)}$ **quando** $\boldsymbol{h \to h_0}$ (com $h_0 \in [-\infty, +\infty]$) se

$$\lim_{h \to h_0} \frac{g(h)}{h^n} = C$$

para algum $C \in \mathbb{R}$. Desta forma, as funções $g = g(h)$ e h^n têm o mesmo comportamento quando $h \to h_0$.

Diz-se que a função real $g = g(h)$ de variável real h é $\boldsymbol{o(h^n)}$ **quando** $\boldsymbol{h \to h_0}$ (com $h_0 \in [-\infty, +\infty]$) se

$$\lim_{h \to h_0} \frac{g(h)}{h^n} = 0$$

para algum $C \in \mathbb{R}$. Desta forma, a função $g = g(h)$ tem um crescimento mais lento que h^n quando $h \to h_0$.

Exercício 5.12. Indique o valor lógico das seguintes afirmações:

(a) $\sin(x) = O(x)$ quando $x \to 0$;

(b) $\cos(x) - 1 = O(x^2)$ quando $x \to 0$;

(c) $x \ln x = O(x)$ quando $x \to \infty$;

(d) $x \ln x = o(x^2)$ quando $x \to \infty$;

(e) $e^x = O(x^n)$ quando $x \to \infty$, para $n \in \mathbb{N}$.

Resposta.

(a) $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$ logo é verdadeiro;

(b) Temos

$$\lim_{x\to 0} \frac{\cos(x)-1}{x^2} = \frac{0}{0}$$

logo pela Regra de Cauchy

$$\lim_{x\to 0} \frac{\cos(x)-1}{x^2} \overset{\text{R.C.}}{=} \lim_{x\to 0} \frac{-\sin(x)}{2x} = -\frac{1}{2},$$

logo é verdadeiro.

(c) $\lim_{x\to\infty} \frac{x\ln x}{x} = \infty$ logo é falso;

(d) Temos

$$\lim_{x\to\infty} \frac{x\ln x}{x^2} = \lim_{x\to\infty} \frac{\ln x}{x} = \frac{\infty}{\infty}$$

logo pela Regra de Cauchy

$$\lim_{x\to\infty} \frac{\ln x}{x} \overset{\text{R.C.}}{=} \lim_{x\to\infty} \frac{1}{x^2} = 0$$

logo é verdadeiro.

(e) Temos

$$\lim_{x\to\infty} \frac{e^x}{x^n} = \frac{\infty}{\infty}$$

logo pela Regra de Cauchy aplicada n vezes sucessivas

$$\lim_{x\to\infty} \frac{e^x}{x^n} \overset{\text{R.C.}}{=} \lim_{x\to\infty} \frac{e^x}{n\,x^{n-1}} \overset{\text{R.C.}}{=} \cdots \overset{\text{R.C.}}{=} \lim_{x\to\infty} \frac{e^x}{n!} = \infty$$

logo é falso.

Utilizando a notação da definição 5.11, podemos escrever a fórmula de Taylor da seguinte forma, considerando $x = x_0 + h$ em que h é uma perturbação pequena, ou seja, $h \approx 0$.

Teorema 5.13 (Fórmula de Taylor).
Seja $f : [a,b] \to \mathbb{R}$ uma função real de variável real $(n+1)$-vezes diferenciável. Então, a função f pode ser representada em torno do ponto $x_0 \in]a,b[$ por

$$f(x_0+h) = P_f^{(n)}(x_0+h) + O(h^{n+1}) \tag{5.4}$$

quando $h \to 0$, em que $P_f^{(n)}$ é o polinómio de Taylor (5.1) em torno de x_0.

Demonstração. Basta tomar $x = x_0 + h$ para verificar que

$$\lim_{h\to 0} \frac{R_f^{(n)}(x_0+h)}{h^{n+1}} = \lim_{h\to 0} \frac{f^{(n+1)}(\xi_h)}{(n+1)!} = \frac{f^{(n+1)}(x_0)}{(n+1)!} \in \mathbb{R}$$

pois $\xi_h \in [x_0; x_0 + h]$. □

Mais uma vez o teorema anterior mostra que o polinómio de Taylor em torno de x_0 aproxima a função f em torno desse ponto, ou seja, para h pequeno em módulo. Nesse caso, quanto maior a ordem n do polinómio de Taylor, mais rápida a convergência quando $h \to 0$, e logo melhor a aproximação de $f(x_0 + h)$ para $h \approx 0$.

Exercício 5.14. Em determinado momento, a polícia detetou um carro que passava no kilómetro 100 da autoestrada $A1$ a uma velocidade de $180km/h$ e com uma aceleração de $-20km/h^2$. Determine aproximadamente a velocidade a que o carro se deslocava e o local onde se encontrava 1 minuto antes deste momento.

Resolução. Sendo $y = y(t)$ a posição do carro no instante t (em horas) e $t = 0$ o instante em que a polícia avistou o carro, pela fórmula de Taylor sabemos que

$$y(t) \approx y(0) + y'(0)t + y''(0)\frac{t^2}{2} = 100 + 180t - 10t^2.$$

Assim, no instante $t = -1/60$, o carro encontrava-se aproximadamente no kilómetro

$$y\left(-\frac{1}{60}\right) = 100 - \frac{180}{60} - \frac{10}{60^2} \approx 96.9972$$

da autoestrada, a uma velocidade

$$y'\left(-\frac{1}{60}\right) = 180 + \frac{20}{60} \approx 180.333\text{km/h}.$$

5.2 Interpolação Polinomial

Nesta secção, vamos supor que conhecemos o valor da função real $f : [a, b] \to \mathbb{R}$ em $n + 1$ pontos distintos $x_0, x_1, \ldots, x_n$ no intervalo $[a, b]$, isto é, conhecemos $f_0, f_1, \ldots, f_n \in \mathbb{R}$ tais que

$$f(x_0) = f_0, \quad f(x_1) = f_1, \quad \ldots, \quad f(x_n) = f_n.$$

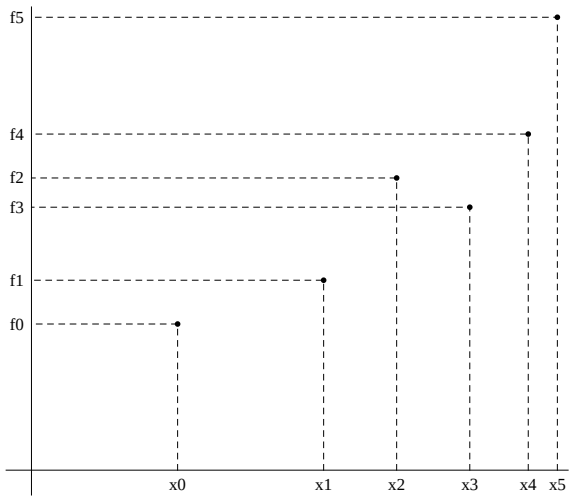


Figura 5.1: *Pontos conhecidos.*

Nota 5.15. Salvo nota em contrário, vamos considerar ao longo desta secção que

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Pretendemos aproximar a função f desconhecida com base nesta informação. Para já relembramos o resultado seguinte, que nos será útil mais adiante. Recordamos então o Teorema Fundamental da Álgebra.

Teorema 5.16 (Teorema Fundamental da Álgebra[1]).
Um polinómio de grau $n \geq 1$ e não-nulo tem exatamente n raízes em $\mathbb{C}$ (contadas com a sua multiplicidade).

Demonstração. A prova sai fora do âmbito deste texto, mas pode ser encontrado em qualquer livro de teoria de funções, como por exemplo [8, 11]. □

O objetivo da interpolação polinomial é encontrar um polinómio p de ordem apropriada que satisfaça as condições dadas, ou seja,

$$p(x_i) = f_i, \quad i = 0, 1, \ldots n.$$

Temos então o seguinte resultado, que garante existência e unicidade desse polinómio.

[1]Como nota histórica convém referir que o problema de determinação de raízes de polinómios foi estudados desde meados do sec. XVIII, contando com a contribuição de grandes matemáticos como Euler, Cauchy, Gauss e Weierstrass, entre outros.

Teorema 5.17 (Polinómio interpolador de Lagrange).

Dados $n+1$ pontos $x_0, x_1, \ldots x_n$ e $n+1$ valores $f_0, f_1, \ldots, f_n$ da função f nesses pontos, existe um único polinómio p_n de grau menor ou igual a n tal que

$$p_n(x_i) = f_i, \quad i = 0, 1, \ldots n. \tag{5.5}$$

A este polinómio chama-se ***polinómio interpolador de f nos pontos $x_0, x_1, \ldots, x_n$.*** *Na representação de Lagrange, este polinómio pode ser dado por*

$$p_n(x) = \sum_{k=0}^{n} f_k l_k(x) \tag{5.6}$$

em que os polinómios l_k de ordem n são dados por

$$l_k(x) := \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}.$$

Nota 5.18. O simbolo $\prod_{i=0}^{n} q_i$ representa o produto das parcelas q_i, com $i = 0, 1, \ldots, n$.

Demonstração. Para demonstrarmos o resultado, temos de mostrar que:

(a) existe um polinómio de grau menor ou igual a n que satisfaz (5.5);

(b) esse polinómio é único.

Dividimos a demonstração nesses dois passos:

(a) Consideramos o polinómio de Lagrange (5.6). É fácil mostrar que os $l_k = l_k(x)$ são polinómios de grau menor ou igual a n e que verificam

$$l_k(x_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases}$$

Assim, substituindo em (5.6), o polinómio de Lagrange é um polinómio de grau menor ou igual a n que satisfaz (5.5).

(b) Suponhamos que existem q_n e r_n dois polinómios distintos de grau menor ou igual a n que satisfazem (5.5). Definimos então o polinómio $p_n = q_n - r_n$ não-nulo que é claramente um polinómio de grau menor ou igual a n e que satisfaz

$$p_n(x_i) = q_n(x_i) - r_n(x_i) = f_i - f_i = 0, \quad i = 0, 1, \ldots, n.$$

Assim p_n é um polinómio não-nulo de grau n com $n+1$ zeros, o que é uma contradição com o Teorema Fundamental da Álgebra 5.16. Assim, não existem dois polinómios distintos de grau menor ou igual que n que satisfaçam (5.5).

□

Exercício 5.19. Conhece-se o valor de f de acordo com a tabela

x_i	0	1	3
f_i	1	2	1

Determine o respetivo polinómio interpolador.

Resposta.

Temos três nós de interpolação

$$x_0 = 0, \quad x_1 = 1, \quad x_2 = 3,$$

e os respetivos valores de f

$$f_0 = 1, \quad f_1 = 2, \quad f_2 = 1,$$

logo procuramos um polinómio de grau menor ou igual a 2. Na representação de Lagrange, temos

$$\begin{aligned} p_2(x) &= \sum_{k=0}^{2} f_k l_k(x) \\ &= f_0 l_0(x) + f_1 l_1(x) + f_2 l_2(x) \\ &= l_0(x) + 2l_1(x) + l_2(x). \end{aligned}$$

Assim, temos

$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{x^2}{3} - \frac{4x}{3} + 1$$

$$l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = -\frac{x^2}{2} + \frac{3x}{2}$$

$$l_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{x^2}{6} - \frac{x}{6}$$

pelo que o polinómio interpolador é dado por

$$p_2(x) = -\frac{x^2}{2} + \frac{3x}{2} + 1.$$

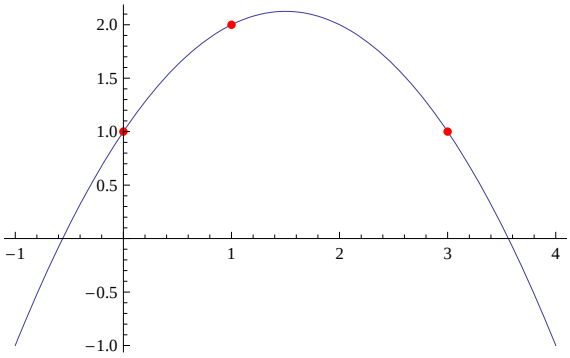


Figura 5.2: *Polinómio interpolador do exercício 5.19.*

Exercício 5.20. Conhece-se o valor de f de acordo com a tabela

x_i	-1	0	2	3
f_i	1	-2	0	-1

.

Determine o respetivo polinómio interpolador de grau 2 que interpola nos 3 primeiros pontos e o polinómio interpolador de grau 3 que interpola em todos os pontos tabelados.

Resposta.

Para os três primeiros nós de interpolação

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 2,$$

e os respetivos valores de f

$$f_0 = 1, \quad f_1 = -2, \quad f_2 = 0,$$

temos pela representação de Lagrange que o polinómio de grau menor ou igual a 2 é dado por

$$\begin{aligned} p_2(x) &= \sum_{k=0}^{2} f_k l_k(x) \\ &= f_0 l_0(x) + f_1 l_1(x) + f_2 l_2(x) \\ &= l_0(x) - 2l_1(x). \end{aligned}$$

Assim, temos

$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{x^2}{3} - \frac{2x}{3}$$

$$l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = -\frac{x^2}{2} + \frac{x}{2} + 1$$

pelo que o polinómio interpolador é dado por

$$p_2(x) = \frac{4x^2}{3} - \frac{5x}{3} - 2.$$

Considerando os quatro nós de interpolação

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 2, \quad x_3 = 3,$$

e os respetivos valores de f

$$f_0 = 1, \quad f_1 = -2, \quad f_2 = 0, \quad f_3 = -1,$$

temos pela representação de Lagrange que o polinómio de grau menor ou igual a 3 é dado por

$$\begin{aligned} p_3(x) &= \sum_{k=0}^{3} f_k l_k(x) \\ &= l_0(x) - 2l_1(x) - l_3(x). \end{aligned}$$

Note-se que temos de recalcular os polinómios l_k, obtendo

$$l_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = -\frac{x^3}{12} + \frac{5x^2}{12} - \frac{x}{2}$$

$$l_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{x^3}{6} - \frac{2x^2}{3} + \frac{x}{6} + 1$$

$$l_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{x^3}{12} - \frac{x^2}{12} - \frac{x}{6}$$

pelo que o polinómio interpolador é dado por

$$p_3(x) = -\frac{x^3}{2} + \frac{11x^2}{6} - \frac{2x}{3} - 2.$$

O exemplo anterior revela uma das grandes desvantagens da representação de Lagrange, uma vez que quando queremos obter um polinómio de ordem superior (acrescentado um ou mais nós), de nada nos serve o polinómio interpolador já calculado. De facto, os polinómios l_k têm de ser recalculados quando se adicionam nós de interpolação. Para evitar este cálculo, vamos introduzir outra forma de calcular o polinómio interpolador. Para isso precisamos da definição seguinte.

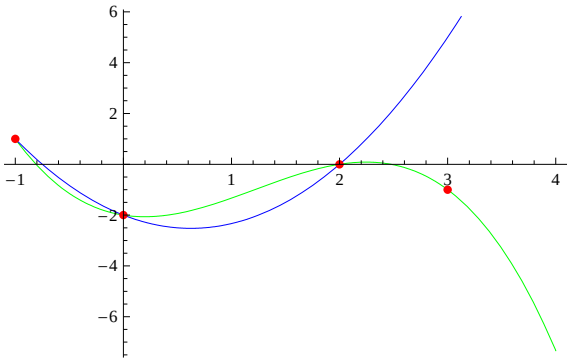


Figura 5.3: *Polinómios interpoladores p_2 (a azul) e p_3 (a verde) do exercício 5.20.*

Definição 5.21 (Diferenças divididas).

Sejam $x_0, x_1, \ldots, x_n$ nós no intervalo $[a, b]$ onde se conhecem os respetivos valores $f_0, f_1, \ldots, f_n$ da função real f.

Chama-se **diferença dividida de primeira ordem no nó x_i** e denota-se por $D_i^1 f$ ou $f[x_i, x_{i+1}]$ à razão

$$f[x_i, x_{i+1}] \equiv D_i^1 f := \frac{f_{i+1} - f_i}{x_{i+1} - x_i}.$$

Por recursão, chama-se **diferença dividida de ordem $k \geq 2$ no nó x_i** e denota-se por $D_i^k f$ ou $f[x_i, x_{i+1}, \ldots, x_{i+k}]$ à razão

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] \equiv D_i^k f := \frac{D_{i+1}^{k-1} f - D_i^{k-1} f}{x_{i+k} - x_i}.$$

Devido à sua definição por recursão, para calcular as diferenças divididas geralmente constroem-se tabelas de diferenças divididas. Vejamos um exemplo.

Exemplo 5.22. Pretendemos determinar as diferenças divididas correspondentes aos nós

x_i	-1	0	2	3
f_i	1	-2	0	-1

.

Temos então a seguinte tabela de diferenças divididas:

x_i	f_i	$D_i^1 f$	$D_i^2 f$	$D_i^3 f$
-1	1	$\frac{-2-1}{0-(-1)} = -3$	$\frac{1-(-3)}{2-(-1)} = \frac{4}{3}$	$\frac{-2/3-(4/3)}{3-(-1)} = -\frac{1}{2}$
0	-2	$\frac{0-(-2)}{2-0} = 1$	$\frac{-1-1}{3-0} = -\frac{2}{3}$	
2	0	$\frac{-1-0}{3-2} = -1$		
3	-1			

Na tabela temos representadas todas as diferenças divididas correspondentes aos nós dados.

Exercício Octave 5.23. Escreva uma função de Octave `DifDiv(lx,lf)` que dados os nós de abcissas `lx` e ordenada `ly` determine a matriz das diferenças divididas.

Resposta.

No ficheiro `DifDiv.m` escrevemos as seguintes linhas de código:

```
function MatDD = DifDiv(lx,lf)
% Calcula diferencas divididas
% INPUT: [lx,lf] - lista de nós
% OUTPUT: MatDD - Diferenças divididas
if size(lx) == size(lf)
  ldif = zeros(size(lx));
  MatDD = NaN(max(size(lx)));
  MatDD(:,1) = lf;
  for ii = 1:  max(size(lx))
    for jj = 1 :  max(size(lx))-ii
      MatDD(jj,ii+1) = (MatDD(jj+1,ii)- MatDD(jj,ii))/...
         (lx(jj+ii)-lx(jj));
    end
  end
else
  disp('Dimensoes de lx e lf nao correspondem!')
end
return
```

As diferenças divididas permitem também obter o polinómio interpolador. Em relação à representação de Lagrange, têm a vantagem de permitir reaproveitar o polinómio de ordem n para obter o polinómio de ordem $n+1$. Além disso permitem facilmente ser implementados numa linguagem de cálculo científico como o Octave. Temos então o seguinte resultado.

Teorema 5.24 (Representação de Newton do polinómio interpolador).

Sejam $x_0, x_1, \ldots x_n$, $n+1$ pontos do intervalo $[a,b]$, sejam $f_0, f_1, \ldots, f_n$ os valores da função f nesses pontos e sejam $D_i^k f$ as respetivas diferenças divididas.

Então o polinómio interpolador p_n de grau menor ou igual a n tal que

$$p_n(x_i) = f_i, \quad i = 0, 1, \ldots n.$$

pode ser dado na representação de Newton por

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} D_0^k f \,(x - x_0)\,(x - x_1)\ldots(x - x_{k-1})\,. \tag{5.7}$$

Demonstração. Indicamos uma forma de fazer esta prova por indução no grau do polinómio n. Como o polinómio interpolador é único e o polinómio interpolador de grau 0 é claramente dado por

$$p_0(x) = f(x_0),$$

para demonstrar (5.7), basta mostrar que o polinómio

$$q_n(x) = p_n(x) - p_{n-1}(x)$$

de grau menor ou igual a n é dado por

$$q_n(x) = D_0^n f \,(x - x_0)\,(x - x_1)\ldots(x - x_{n-1})\,.$$

Como p_n e p_{n-1} são polinómio interpoladores, temos para os dois que

$$p_n(x_i) = f_i, \quad p_{n-1}(x_i) = f_i, \qquad i = 0, 1, \ldots, n-1$$

logo o polinómio q satisfaz

$$q_n(x_i) = 0, \qquad i = 0, 1, \ldots, n-1.$$

Assim, como um polinómio de grau n tem exatamente n raízes, temos que

$$q_n(x) = C_n\,(x - x_0)\,(x - x_1)\ldots(x - x_{n-1})$$

faltando apenas mostrar que C_n, que é o coeficiente do termo x^n é dado por

$$C_n = D_0^n f.$$

Esta prova sai diretamente por indução em n. Ver [5]. □

Exercício 5.25. Determine o polinómio interpolador de grau 2 e 3 pela representação de Newton, considerando os primeiros nós da tabela

x_i	-1	0	2	3
f_i	1	-2	0	-1

.

Resolução. A tabela das diferenças divididas já foi construída no exemplo 5.22. Assim

$$\begin{aligned} p_2(x) &= f(x_0) + D_0^1 f\,(x - x_0) + D_0^2 f\,(x - x_0)(x - x_1) \\ &= 1 - 3(x+1) + \frac{4}{3}x(x+1) \\ &= \frac{4x^2}{3} - \frac{5x}{3} - 2. \end{aligned}$$

Da mesma forma temos que

$$\begin{aligned} p_3(x) &= p_2(x) + D_0^3 f\,(x - x_0)(x - x_1)(x - x_2) \\ &= p_2(x) - \frac{1}{2}x(x+1)(x-2) \\ &= -\frac{x^3}{2} + \frac{11x^2}{6} - \frac{2x}{3} - 2. \end{aligned}$$

Tal como esperado, o resultado coincide com o obtido no exercício 5.20.

Exercício Octave 5.26 (Polinómio interpolador).

Utilize a função `polyfit(lx,ly,n)` (que dadas as listas das abcissas `lx` e das ordenadas `ly` de $n+1$ nós de interpolação determina os coeficientes do polinómio interpolador) para determinar a resposta do exercício anterior. Utilize a função `polyval(coef,x)` (que determina o valor do polinómio de coeficientes `coef` em `x`) para determinar o valores dos polinómios interpoladores em $x = 1$.

Resposta.

Temos $p_2(1) = -7/3, \quad p_3(1) = -4/3.$

Exercício 5.27. Determine os polinómios interpoladores nos nós tabelados:

(a)

x_i	0	1	2	3
f_i	1	2	5	10

;

(b)
x_i	-2	-0.5	0.5	2
f_i	0.8	0.2	0.3	-0.1

;

(c)
x_i	0.2	0.4	0.7
f_i	0.21	0.34	-0.57

;

(d)
x_i	0.1	0.2	0.3	0.4
f_i	1.1	1.9	3.0	3.9

;

Resposta.

(a) $p_3(x) = x^2 + 1$;

(b) $p_3(x) = -0.0866667x^3 + 0.0266667x^2 + 0.121667x + 0.243333$;

(c) $p_2(x) = -7.36667x^2 + 5.07x - 0.509333$;

(d) $p_3(x) = -83.3333x^3 + 65x^2 - 5.66667x + 1.1$

Considerando que o polinómio interpolador p_n é uma aproximação da função desconhecida f é importante conseguir estabelecer uma estimativa de erro para a aproximação. Assim queremos estimar o erro absoluto

$$e_x = |f(x) - p_n(x)|$$

para qualquer ponto x no intervalo $[a, b]$. Claro que se x coincidir com um dos nós x_i o erro é zero. Tomamos então x como sendo um novo nó e p_{n+1} o polinómio de ordem $n + 1$ obtido, isto é,

$$p_{n+1}(y) = p_n(y) + f[x_0, x_1, \ldots, x_n, x](y - x_0)(y - x_1)\ldots(y - x_n).$$

Como $p_{n+1}(x) = f(x)$, para $y = x$ temos o erro

$$e_x = |f(x) - p_n(x)| = |f[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1)\ldots(x - x_n)|\,. \qquad (5.8)$$

A fórmula anterior não ajuda a estimar o erro, uma vez que o valor de $f(x)$ é desconhecido e logo não é possível calcular a diferença dividida $f[x_0, x_1, \ldots, x_n, x]$. Para resolver esta questão temos o teorema seguinte.

Teorema 5.28.
Seja $f : [a, b] \to \mathbb{R}$ *uma função* $n + 1$ *vezes diferenciável e sejam*

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Então, existe um $\xi \in [a, b]$ *tal que*

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

O teorema anterior, juntamente com a fórmula do erro (5.8) permite obter a seguinte estimativa do erro.

Corolário 5.29 (Estimativa de erro de interpolação).
Seja $f : [a, b] \to \mathbb{R}$ uma função $n+1$ vezes diferenciável e seja $p_n = p_n(x)$ o polinómio interpolador de grau menor ou igual a n nos pontos

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Então, para cada $x \in [a, b]$ existe um $\xi \in [a, b]$ tal que

$$|f(x) - p_n(x)| = \left| \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\dots(x - x_n) \right|.$$

Por outras palavras, o erro pode ser majorado para $x \in [a, b]$ por

$$|f(x) - p_n(x)| \leq \frac{\max\limits_{\xi\in[a,b]} \left|f^{(n+1)}(\xi)\right|}{(n+1)!} |(x - x_0)(x - x_1)\dots(x - x_n)|. \tag{5.9}$$

O teorema anterior diz-nos que para estabelecer uma estimativa de erro é necessário ter conhecimento adicional sobre a função, nomeadamente sobre a magnitude das suas derivadas. Este conhecimento pode vir muitas vezes do conhecimento empírico que se tem dos problemas.

Exercício 5.30. Determine uma aproximação para $f(0.7)$, tomando os valores da tabela seguinte:

x_i	0	0.2	0.6	1
f_i	1.1	1.2	0.9	0.6

;

Determine uma estimativa de erro, sabendo que o valor absoluto da quarta derivada de f é menor que 3.

Resposta. Obtemos o polinómio interpolador

$$p_3(x) = 2.08333x^3 - 3.75x^2 + 1.16667x + 1.1$$

pelo que a aproximação é dada por

$$f(0.7) \approx p_3(0.7) = 0.79375.$$

De (5.9) temos

$$\begin{aligned} |f(0.7) - p_3(0.7)| &\leq \frac{\max\limits_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right|}{4!} |(0.7 - 0)(0.7 - 0.2)(0.7 - 0.6)(0.7 - 1)| \\ &= 0.0013125 \end{aligned}$$

logo os dois primeiros dígitos não nulos da aproximação estão corretos.

Exercício 5.31. Considere a função $f(x) = \sin(x)$.

(a) Determine o polinómio interpolador de f nos pontos

$$x_0 = 0,\ x_1 = \frac{\pi}{4},\ x_2 = \frac{\pi}{2}.$$

(b) Determine uma aproximação de $f(0.3)$ pelo polinómio interpolador.

(c) Determine uma estimativa de erro da aproximação anterior e compare-a com o erro efetivo.

Resposta. (a) $p_2(x) = 1.16401x - 0.335749x^2$

(b) $f(0.3) \approx p_2(0.3) = 0.318986$.

(c) Uma vez que

$$|f'''(x)| = |-\cos x| \leq 1,$$

a estimativa de erro é dada por

$$|f(0.3) - p_2(0.3)| \leq 0.0308421$$

enquanto que o erro efetivo é

$$|f(0.3) - p_2(0.3)| = 0.0234663.$$

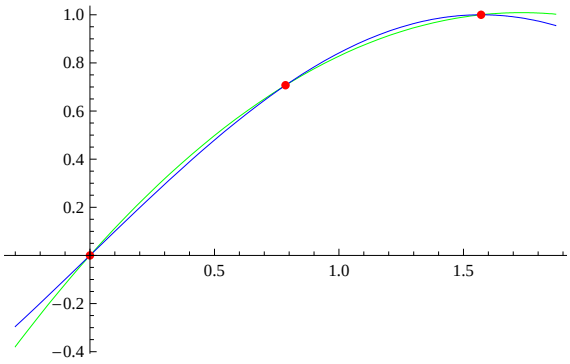


Figura 5.4: *Polinómio interpolador do exercício 5.31 (a verde) e a função f (a azul).*

Exercício 5.32. Considere a função $f(x) = e^x$.

(a) Determine o polinómio interpolador de f nos pontos

$$x_0 = 0,\ x_1 = 0.5,\ x_2 = 1,\ x_3 = 2$$

(b) Determine uma aproximação de $f(1.5)$ pelo polinómio interpolador.

(c) Determine uma estimativa de erro da aproximação anterior e compare-a com o erro efetivo.

Resposta. (a) $p_3(x) = 0.423045x^3 + 0.207111x^2 + 1.08813x + 1$.

(b) $f(1.5) \approx p_3(1.5) = 4.52597$.

(c) Uma vez que

$$f^{(4)}(x) = e^x,$$

é uma função crescente, temos

$$\max_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right| = e^2 \approx 7.38906.$$

Assim, a estimativa de erro é dada por

$$|f(1.5) - p_3(1.5)| \leq 0.115454$$

enquanto que o erro efetivo é

$$|f(1.5) - p_3(1.5)| = 0.0442764.$$

É conhecido que se o número de nós de interpolação (e consequentemente o número do grau do polinómio) aumentar, a aproximação fica instável. Isto acontece particularmente quando se tem nós igualmente espaçados, sendo o fenómeno conhecido como **fenómeno de Runge**. Por exemplo, se interpolarmos a função $f(x) = \cos(x)\sin(x)$ no intervalo $[0, 6\pi]$, o erro nos extremos do intervalo parece crescer, conforme se ilustra na figura 5.6). Isto deve-se ao facto de na estimativa de erro (5.9) o termo

$$|(x - x_0)(x - x_1)\ldots(x - x_n)|$$

assumir valores muito mais altos nos extremos do intervalo $[a, b]$ em comparação com os valores centrais do intervalo.

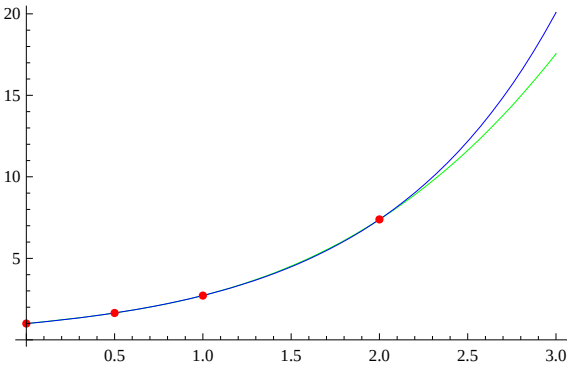


Figura 5.5: *Polinómio interpolador do exercício 5.32 (a verde) e a função f (a azul).*

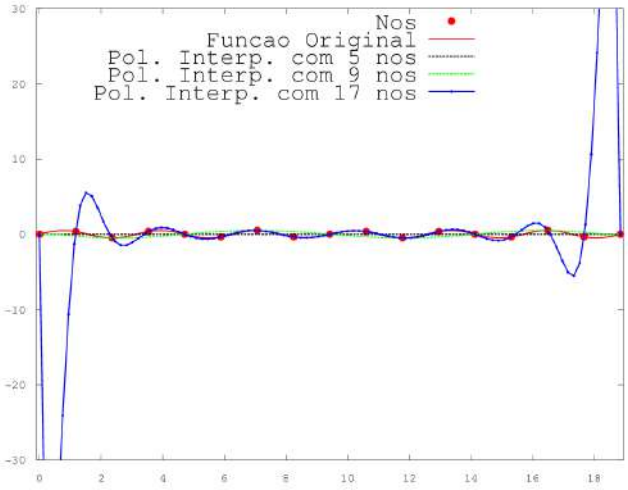


Figura 5.6: *Ilustração do Fenómeno de Runge: instabilidade nos extremos do intervalo para polinómios interpoladores de grau elevado, usando nós igualmente espaçados.*

Nesta perspetiva é importante encontrar uma solução interpoladora que seja estável com o aumentar do número de nós de interpolação. Uma hipótese que pode controlar essa propagar de erro é utilizar nós não igualmente espaçados, como por exemplo os nós de Chebyshev (ver [5, 10]). No entanto, em muitas aplicações os dados são recolhidos em nós igualmente espaçados, pelo que temos de encontrar outra alternativa. É nessa perspetiva que são introduzidos os splines interpoladores.

5.3 Interpolação por splines

A interpolação por splines consiste em considerar uma aproximação global suave num intervalo $[a, b]$ conhecidos os valores da função f nos nós

$$a = x_0 < x_1 < \cdots < x_n = b$$

em que em cada subintervalo $[x_k, x_{k+1}]$, $k = 0, 1, \ldots, n$ a função é um polinómio. Em particular, a regra é que se seccionalmente o polinómio é de grau p então globalmente o spline será $p-1$ vezes diferenciável, isto é, pertence a $C^{p-1}([a, b])$.

Por uma questão de simplicidade, começamos pelo caso linear. Assim exigimos ao spline linear que seja um polinómio de grau 1 em cada subintervalo e que seja contínuo em todo o intervalo. Por conveniência vamos definir o espaçamento entre os nós por

$$h_k = x_{k+1} - x_k.$$

Definição 5.33 (Spline linear).

Seja $f : [a, b] \to \mathbb{R}$ uma função em $C^2([a, b])$. Define-se como **spline linear** de f nos nós

$$a = x_0 < x_1 < \cdots < x_n = b$$

a função S_1 definida por ramos

$$S_1(x) = \begin{cases} S_{1,0}(x), & x \in [x_0, x_1] \\ S_{1,1}(x), & x \in [x_1, x_2] \\ \vdots & \\ S_{1,n-1}(x), & x \in [x_{n-1}, x_n] \end{cases}$$

que em cada intervalo $[x_k, x_k + 1]$, $k = 0, 1, \ldots, n-1$ é definida pelo polinómio interpolador $S_{1,k}$ de grau 1 nos nós x_k e x_{k+1}, ou seja, temos a partir de (5.6) que

$$S_{1,k}(x) = -f_k \frac{x - x_{k+1}}{h_k} + f_{k+1} \frac{x - x_k}{h_k}, \quad x \in [x_k, x_k + 1],\ k = 0, 1, \ldots, n-1,$$

com $h_k = x_{k+1} - x_k$. De forma semelhante, temos a estimativa de erro dada para $x \in [x_k, x_k + 1]$ com $k = 0, 1, \ldots, n-1$ por

$$|f(x) - S_{1,k}(x)| \leq \frac{\max\limits_{\xi \in [x_k, x_{k+1}]} |f''(\xi)|}{2} |(x - x_k)(x - x_{k+1})| \tag{5.10}$$

e a estimativa global para $x \in [a, b]$ dada por

$$|f(x) - S_1(x)| \leq \frac{\max\limits_{\xi \in [a,b]} |f''(\xi)|}{8} \max_{k=0,1,\ldots n-1} h_k^2. \tag{5.11}$$

Demonstração. A estimativa de erro (5.10) sai diretamente da estimativa (5.9), aplicada a cada intervalo $[x_k, x_{k+1}]$. A estimativa (5.11) sai da anterior, notando que o máximo de

$$g(x) = |(x - x_k)(x - x_{k+1})|$$

é obtido no ponto médio do intervalo $\dfrac{x_k + x_{k+1}}{2}$ com valor

$$g\left(\frac{x_k + x_{k+1}}{2}\right) = \frac{(x_{k+1} - x_k)^2}{4} = \frac{h_k^2}{4}.$$

□

Exercício Octave 5.34.

Faça o gráfico do spline linear que aproxima a função $f(x) = x^2 \sin(3x)$ no intervalo $[0, 20]$ nos nós $x_k = k/2,\ k = 0, 1, \ldots 20$, utilizando a função

```
interp1(lx,ly,lxi)
```

em que `lx` é a lista dos nós, `ly` é a lista das imagens de f e `lxi` é a lista de nós onde se quer saber a imagem do spline linear.

Resolução.

Executando a lista de comandos seguinte, obtém-se a figura 5.7:

```
lx = 0:0.5:10;
ly = lx.^ 2.*sin(3*lx);
lxp = 0:0.05:10;
lyp = lxp.^2.*sin(3*lxp);
lyS = interp1(lx,ly,lxp);
plot(lxp,lyp,'r-',lxp,lyS,'b-',lx,ly,'bo')
legend('Original','Spline linear','Nos de interpolacao')
```

Note-se que no caso linear, o spline pode ser interpretado como uma aproximação por um polinómio segmentado. Nesse caso o que se faz é agrupar os $n+1$ nós em grupos de $p+1$ nós e em cada um deles aproximar por um polinómio de grau p, como na secção anterior. Assim, no caso em que n é múltiplo de p, no intervalo $[x_0, x_p]$ teríamos um polinómio interpolador de grau p nos $p+1$ nós deste intervalo, no intervalo $[x_p, x_{2p}]$ teríamos um segundo polinómio interpolador de grau p nos $p+1$ nós deste intervalo, continuando com o processo até ao último intervalo $[x_{n-p}, x_n]$. Desta forma construímos uma aproximação seccionalmente polinomial, continua em [a,b] mas que não diferenciável nos pontos de colagem dos intervalos $x_p, x_{2p}, \ldots, x_{n-p}$.

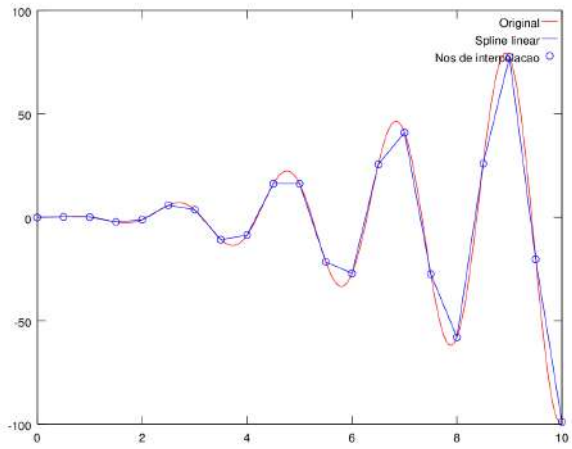


Figura 5.7: *Comparação entre spline linear e função original no caso do exercício 5.34.*

O contexto dos splines de ordem superior à linear é exatamente impor condições de diferenciabilidade em todos os nós. Em vez de agrupar os nós em conjuntos de $p+1$ nós e interpolar por um polinómio de ordem p, o objetivo é considerar um polinómio de ordem p em cada intervalo $[x_k, x_{k+1}]$, $k = 0, 1, \ldots n-1$, entre os $n+1$ nós disponíveis. Para $p = 1$, ou seja, no caso linear que acabámos de ver, as condições de interpolação determinam o polinómio linear. No caso de um grau de polinómio $p \geq 2$, além das condições de interpolação, ficamos com graus de liberdade para impor condições de diferenciabilidade nos nós.

Passamos então a apresentar o spline cúbico S_3. Neste caso, em cada intervalo $[x_k, x_{k+1}]$, $k = 0, 1, \ldots n-1$, o spline será um polinómio cúbico, ou seja, temos de forma semelhante ao caso linear que S_3 é definida por ramos como

$$S_3(x) = \begin{cases} S_{3,0}(x), & x \in [x_0, x_1] \\ S_{3,1}(x), & x \in [x_1, x_2] \\ \vdots & \\ S_{3,n-1}(x), & x \in [x_{n-1}, x_n] \end{cases}$$

em que em cada ramo $[x_k, x_{k+1}]$, $k = 0, 1, \ldots n-1$, temos um polinómio de grau 3 da forma

$$S_{3,k}(x) = a_{0,k} + a_{1,k}x + a_{2,k}x^2 + a_{3,k}x^3.$$

Desta forma, para determinar o spline cúbico temos $4n$ incógnitas a determinar, nomeadamente os quatro coeficientes de cada um dos n polinómios $S_{3,k}$ para $k = 0, 1, \ldots n$. Para determinar unicamente este spline cúbico, temos então de

impor $4n$ condições. Temos $2n$ condições de interpolação

$$S_{3,k}(x_k) = f_k, \quad S_{3,k}(x_{k+1}) = f_{k+1}, \quad k = 0, 1, \ldots, n-1. \tag{5.12}$$

Temos portanto, neste momento $2n$ graus de liberdade no nosso sistema. Assim, impomos $2(n-1)$ condições para que S_3 seja uma duas vezes diferenciável em $[a, b]$, nomeadamente nos pontos x_k, $k = 1, 2, \ldots n-1$ interiores dadas por

$$S'_{3,k-1}(x_k) = S'_{3,k}(x_k) \tag{5.13}$$

$$S''_{3,k-1}(x_k) = S''_{3,k}(x_k), \tag{5.14}$$

para $k = 1, 2, \ldots, n-1$.

Finalmente ficam a faltar duas condições para eliminar os dois graus de liberdade existentes, nomeadamente no primeiro e último intervalo. Quando não há mais informação conhecida, opta-se geralmente um das duas seguintes. Geralmente o software de cálculo científico como o Octave tem implementado o **spline *not-a-knot***, que consiste em adicionar condições de continuidade na terceira derivada nos segundo e penúltimo nós, ou seja,

$$S'''_{3,0}(x_1) = S'''_{3,1}(x_1) \quad \text{e} \quad S'''_{3,n-2}(x_{n-1}) = S'''_{3,n-1}(x_{n-1}). \tag{5.15}$$

Outra hipótese é assumir que as segundas derivadas se anulam nos extremos do intervalo, ou seja,

$$S''_{3,0}(x_0) = S''_{3,n-1}(x_n) = 0, \tag{5.16}$$

o que se traduz no **spline cúbico natural**. Este é geralmente considerado na literatura, por permitir mais facilmente obter majorantes de erro.

Caso se tenha informação adicional, outras condições podem ser impostas em vez das anteriores. Por exemplo se a função a interpolar for periódica, impõe-se

$$S'_{3,0}(x_0) = S'_{3,n-1}(x_n), \quad S''_{3,0}(x_0) = S''_{3,n-1}(x_n),$$

ou caso seja conhecida a derivada no primeiro e último nó consideram-se as condições

$$S'_{3,0}(x_0) = f'(x_0), \quad S'_{3,n-1}(x_n) = f'(x_n),$$

que em ambos os casos se traduzem também em equações lineares nos coeficientes dos polinómios.

A forma de cálculo do spline cúbico é engenhosa. Em vez de ser resolver o sistema $4n \times 4n$, reduz-se a resolução para um sistema $(n+1) \times (n+1)$, em que as incógnitas são os valores das segundas derivadas nos nós definidas por

$$s_k := S''_3(x_k), \quad k = 0, 1, \ldots n.$$

Por definição, é evidente que a segunda derivada $S_3''(x)$ do spline cúbico é contínua em $[a, b]$ e que é um polinómio linear em cada intervalo $[x_k, x_{k+1}]$ para $k = 0, 1, \ldots, n-1$. Por outras palavras, a função $S_3''(x)$ tem a forma de um spline linear, pelo que

$$S_{3,k}''(x) = -s_k \frac{x - x_{k+1}}{h_k} + s_{k+1} \frac{x - x_k}{h_k}, \quad k = 0, 1, \ldots, n-1.$$

Integrando duas vezes em x, temos

$$S_{3,k}(x) = -\frac{s_k}{6} \frac{(x - x_{k+1})^3}{h_k} + \frac{s_{k+1}}{6} \frac{(x - x_k)^3}{h_k} + \alpha x + \beta,$$

que pode ser reescrito como

$$S_{3,k}(x) = -\frac{s_k}{6} \frac{(x - x_{k+1})^3}{h_k} + \frac{s_{k+1}}{6} \frac{(x - x_k)^3}{h_k} + \tilde{\alpha}(x - x_k) + \tilde{\beta}(x - x_{k+1})$$

para $k = 0, 1, \ldots, n-1$. Assim, tomando as condições de interpolação (5.12), temos o sistema

$$\begin{cases} \frac{s_k}{6} h_k^2 - \tilde{\beta} h_k = f_k \\ \frac{s_{k+1}}{6} h_k^2 + \tilde{\alpha} h_k = f_{k+1} \end{cases}$$

logo resolvendo em ordem a $\tilde{\alpha}$ e $\tilde{\beta}$ obtemos

$$\begin{aligned} S_{3,k}(x) = &- \frac{s_k}{6} \frac{(x - x_{k+1})^3}{h_k} + \frac{s_{k+1}}{6} \frac{(x - x_k)^3}{h_k} \\ &+ \left(\frac{f_{k+1}}{h_k} - \frac{s_{k+1} h_k}{6} \right) (x - x_k) - \left(\frac{f_k}{h_k} - \frac{s_k h_k}{6} \right) (x - x_{k+1}) \end{aligned} \tag{5.17}$$

A equação anterior mostra que os $n+1$ valores de s_k definem o spline cúbico. Assim, para determinar os valores vamos usar as condições de continuidade da primeira derivada de S_3 em (5.13), as únicas ainda não utilizadas neste processo. Assim, derivando em x a expressão (5.17) obtemos

$$S_{3,k}'(x) = -\frac{s_k}{2} \frac{(x - x_{k+1})^2}{h_k} + \frac{s_{k+1}}{2} \frac{(x - x_k)^2}{h_k} + \frac{f_{k+1} - f_k}{h_k} + h_k \frac{s_k - s_{k+1}}{6}$$

e impondo as condições (5.13), obtemos

$$\frac{s_k h_{k-1}}{2} + \frac{f_k - f_{k-1}}{h_{k-1}} + h_{k-1} \frac{s_{k-1} - s_k}{6} = -\frac{s_k h_k}{2} + \frac{f_{k+1} - f_k}{h_k} + h_k \frac{s_k - s_{k+1}}{6}$$

para $k = 1, 2, \ldots n-1$, ou seja,

$$s_{k-1}\frac{h_{k-1}}{6} + s_k\frac{h_k + h_{k-1}}{3} + s_{k+1}\frac{h_k}{6} = \frac{f_{k+1} - f_k}{h_k} - \frac{f_k - f_{k-1}}{h_{k-1}}. \tag{5.18}$$

As condições anteriores definem $n-1$ condições do sistema linear, sendo que são necessárias as duas adicionais consoante o spline cúbico considerado. Por exemplo, uma vez que

$$S'''_{3,k}(x) = \frac{s_{k+1} - s_k}{h_k},$$

temos para o spline *not-a-knot* definido por (5.15) as duas condições adicionais

$$\begin{aligned} \frac{s_1 - s_0}{h_0} &= \frac{s_2 - s_1}{h_1}, \\ \frac{s_{n-1} - s_{n-2}}{h_{n-1}} &= \frac{s_n - s_{n-1}}{h_{n-1}} \end{aligned}$$

Por outro lado, tomando o spline cúbico natural que satisfaz (5.16), temos $s_0 = s_n = 0$. Assim, de (5.18), os restantes valores s_k são obtidos por resolução do sistema linear

$$\begin{bmatrix} a_1 & b_1 & 0 & \ldots & 0 \\ b_1 & a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b_{n-3} & a_{n-2} & b_{n-2} \\ 0 & \ldots & 0 & b_{n-2} & a_{n-1} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-2} \\ s_{n-1} \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{n-2} \\ g_{n-1} \end{bmatrix} \tag{5.19}$$

em que

$$\begin{aligned} a_k &= \frac{h_k + h_{k-1}}{3}, \quad k = 1, \ldots n-1, \\ b_k &= \frac{h_k}{6}, \quad k = 1, \ldots n-2, \\ g_k &= \frac{f_{k+1} - f_k}{h_k} - \frac{f_k - f_{k-1}}{h_{k-1}}, \quad k = 1, \ldots n-1. \end{aligned}$$

Chamamos mais uma vez a atenção para ao facto do sistema a resolver não ser da ordem de $4n$ (número de incógnitas do spline cúbico), mas da ordem de n. Em termos de custo computacional, isto é obviamente uma vantagem. Fazemos um resumo do spline cúbico na definição seguinte, à qual adicionamos as estimativas de erro para o spline cúbico.

Definição 5.35 (Spline cúbico).

Seja $f : [a, b] \to \mathbb{R}$ uma função em $C^4([a, b])$. Define-se como **spline cúbico** de f nos nós

$$a = x_0 < x_1 < \cdots < x_n = b$$

a função $S_3 \in C^2([a, b])$ definida por ramos

$$S_3(x) = \begin{cases} S_{3,0}(x), & x \in [x_0, x_1] \\ S_{3,1}(x), & x \in [x_1, x_2] \\ \vdots & \\ S_{3,n-1}(x), & x \in [x_{n-1}, x_n] \end{cases}$$

que em cada intervalo $[x_k, x_k + 1]$, $k = 0, 1, \ldots, n - 1$ é um polinómio $S_{3,k}$ de grau 3. A definição do **spline cúbico natural** que satisfaz (5.16) pode ser feita pela expressão (5.17) em que os elementos s_k são definidos como solução do sistema (5.19).

Temos também as estimativas de erro para a função

$$|f(x) - S_3(x)| \leq \frac{5 \max\limits_{\xi \in [a,b]} \left|f^{(4)}(\xi)\right|}{384} \max_{k=0,1,\ldots n-1} h_k^4 \tag{5.20}$$

e para a derivada

$$|f'(x) - S_3'(x)| \leq \frac{\max\limits_{\xi \in [a,b]} \left|f^{(4)}(\xi)\right|}{24} \max_{k=0,1,\ldots n-1} h_k^3 \tag{5.21}$$

Demonstração. A demonstração das estimativas de erro é demasiado extensa, pelo que remetemos para um livro de análise numérica, como por exemplo [10]. □

Exercício Octave 5.36.

Determine uma função em Octave

```
SplineCubicoNatural(lx,lf,lxp)
```

que dados os vetores linha `lx` e `lf` com os nós e respetivos valores da função f a interpolar, respetivamente, e a lista de pontos `lxp`, determine a lista `lfp` dos valores do spline cúbico natural que nesses pontos.

Resolução.

Escrevemos no ficheiro `SplineCubicoNatural.m` a lista de comandos seguinte:

```
function [lfp]=SplineCubicoNatural(lx,lf,lxp)
n=length(lx);
h=lx(2:end)-lx(1:end-1);
la=(h(2:end)+h(1:end-1))/3;
lb=h(2:end-1)/6;
lfh=lf(2:end)-lf(1:end-1);
lg=lfh(2:end)./h(2:end)-lfh(1:end-1)./h(1:end-1);
A=spdiags([[lb.';0],la.',[0;lb.']],-1:1,n-2,n-2);
sk=A\(lg.');
sk=[0;sk;0];
coefs=zeros(n-1,4);
coefs(:,1)=(sk(2:end)-sk(1:end-1))./6./(h.');
coefs(:,2)=(sk(1:end-1))./2;
coefs(:,3)=-(sk(1:end-1).*(h.'))./2+((lf(2:end)./h).'...
  -sk(2:end).*h.'/6)-((lf(1:end-1)./h).'-sk(1:end-1).*h.'/6);
coefs(:,4)=(sk(1:end-1).*(h.'.^2))./6+((lf(1:end-1)).'...
  -sk(1:end-1).*(h.'.^2)/6);
CS=mkpp(lx,coefs);
lfp=ppval(CS,lxp);
return
```

De notar que o sistema linear a resolver (5.19) é tridiagonal, pelo que há vantagem (em termos computacionais) de utilizar métodos de resolução de sistemas lineares próprios para este tipo de matrizes. Assim, ao definir a matriz A como tridiagonal (atarvés do comando `spdiags`) estamos a "informar"o Octave de que o sistema a resolver é tridiagonal. Para sistemas muito grandes, métodos iterativos para a resolução de sistemas também poderiam ser utilizados com vantagem computacional, como os apresentados no capítulo 7.1.

Exercício Octave 5.37.

Adaptando o pedido no exercício 5.34, faça os gráficos dos splines cúbicos

- *not-a-knot* que aproxima a função $f(x) = x^2 \sin(3x)$ no intervalo $[0, 10]$ nos nós $x_k = k/2$ para $k = 0, 1, \ldots 20$, utilizando a função

  ```
  interp1(lx,ly,lxp,'spline')
  ```

- natural, que aproxima a mesma função nos mesmo nós, utilizando a função do exercício anterior,

em que `lx` é a lista dos nós, `ly` é a lista das imagens de f e `lxp` é a lista de nós onde se quer saber a imagem do spline linear.

Resolução.

Executando a lista de comandos seguinte, obtém-se a figura 5.8:

```
lx = 0:0.5:10;
ly = lx.^ 2.*sin(3*lx);
lxp = 0:0.05:10;
lyp = lxp.^2.*sin(3*lxp);
lySCNaK = interp1(lx,ly,lxp,'spline');
lySCN = SplineCubicoNatural(lx,ly,lxp);
plot(lxp,lyp,'r-',lxp,lySCNaK,'k-',lxp,lySCN,'b-',lx,ly,'r.')
legend('Original','Spline Cubico not-a-knot','Spline Cubico
natural','Nos de interpolacao')
```

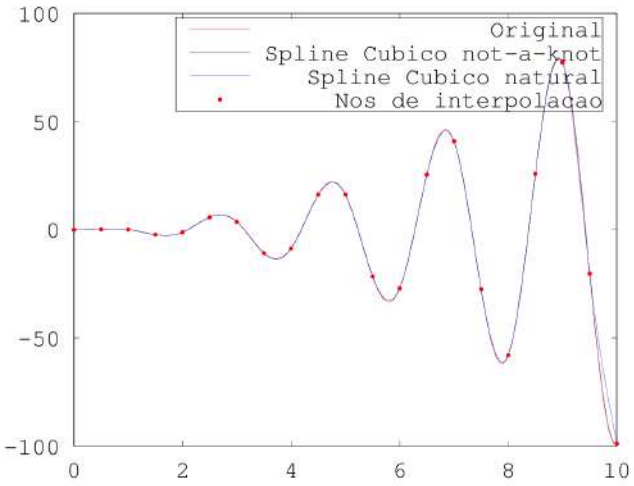


Figura 5.8: *Comparação entres os splines cúbicos* not-a-knot *e natural e a função original no caso do exercício 5.37.*

Quando a função f a aproximar é suave, a aproximação por splines cúbicos é geralmente muito melhor que a aproximação por splines lineares, como se pode ver se compararmos as figuras 5.7 e 5.8 para o mesmo exemplo. De notar também que no exemplo anterior, como a segunda derivada da função interpolanda não se anula no extremo direito, temos uma diferença entre os spline cúbicos *not-a-knot* e natural. Como o spline natural (por definição)impõe que a segunda derivada é nula nos extremos do intervalo, junto ao extremo direito a aproximação não é tão boa por esta via. Esta diferença não é visível no extremo esquerdo

do intervalo, uma vez que nesse extremo a segunda derivada da função interpolanda se anula (conforme é imposto no spline cúbico natural).

Para completar este capítulo sobre aproximação de funções, falta agora apenas ver o caso em que não desejamos que os valores dos nós sejam interpolados. Esse é o caso, por exemplo, quando os dados estão afetados por ruído, ou seja, têm erros associados.

5.4 Aproximação por Mínimos Quadrados

As aproximações de funções até agora consideradas não consideram um dos fatores mais importantes da matemática aplicada: as medições não são em geral exatas, estando afetadas por erros. Desta forma os valores de f nos $n+1$ nós

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

não são exatos, pelo que obrigar a função aproximada $\tilde{f}$ a satisfazer

$$\tilde{f}(x_i) = f_i, \quad i = 0, 1, \ldots n$$

(como no caso da interpolação) pode ser errado, caso a magnitude do erro na medição dos valores f_i seja elevada.

A aproximação por mínimos quadrados parte do princípio que os valores f_i estão afetados de erros e que temos algum conhecimento da função a aproximar. Suponhamos que pretendemos aproximar f por uma função da forma

$$\tilde{f}(x) = a_0\phi_0(x) + a_1\phi_1(x) + \cdots + a_m\phi_m(x)$$

em que as $m+1$ funções reais de variável real

$$\phi_j = \phi_j(x), \quad j = 0, 1, \ldots, m$$

linearmente independentes se denominam funções base. Pretendemos determinar os coeficientes $a_j, j = 0, 1, \ldots m$ que melhor aproximem os pontos (x_i, f_i), $i = 0, 1, \ldots n$ dados. No sentido dos mínimos quadrados, a melhor aproximação é entendida como encontrar os coeficientes $a_j, j = 0, 1, \ldots m$ que minimizem a função de custo

$$J(a_0, a_1, \ldots, a_m) := \sum_{i=0}^{n} \left(\tilde{f}(x_i) - f_i\right)^2, \tag{5.22}$$

ou seja, que minimizem a soma dos quadrados dos erros nos pontos x_i. Ao valor da função anterior chama-se erro no sentido dos mínimos quadrados da aproximação $\tilde{f}$.

Temos então o seguinte resultado.

Teorema 5.38 (Aproximação por mínimos quadrados).

Sejam $x_0, x_1, \ldots x_n$, os $n+1$ nós correspondentes aos valores $f_0, f_1, \ldots, f_n$, medidos de f. Sejam ainda as $m+1$ funções de base (reais)

$$\phi_j = \phi_j(x), \quad j = 0, 1, \ldots, m$$

linearmente independentes, com $m \leq n$. Então, os coeficientes $a_j, j = 0, 1, \ldots m$ correspondentes à melhor aproximação

$$\tilde{f}(x) = a_0\phi_0(x) + a_1\phi_1(x) + \cdots + a_m\phi_m(x)$$

por mínimos quadrados são a solução do sistema linear

$$\underbrace{\begin{bmatrix} (\phi_0,\phi_0) & (\phi_0,\phi_1) & \ldots & (\phi_0,\phi_m) \\ (\phi_1,\phi_0) & (\phi_1,\phi_1) & \ldots & (\phi_1,\phi_m) \\ \vdots & \vdots & \ddots & \vdots \\ (\phi_m,\phi_0) & (\phi_m,\phi_1) & \ldots & (\phi_m,\phi_m) \end{bmatrix}}_{A} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \underbrace{\begin{bmatrix} (\phi_0,f) \\ (\phi_1,f) \\ \vdots \\ (\phi_m,f) \end{bmatrix}}_{b} \tag{5.23}$$

em que o produto interno discreto $(.,.)$ envolvido é dado por

$$(f,g) = \sum_{i=0}^{n} f(x_i)g(x_i).$$

Demonstração. Queremos minimizar a função J em (5.22) em relação aos coeficientes $a_j, j = 0, 1, \ldots m$. A ideia passa por derivar J em cada um dos coeficientes e igualar as derivadas a zero. Assim, temos

$$\begin{aligned}
\frac{\partial J}{\partial a_j}(a_0, \ldots, a_m) &= \frac{\partial}{\partial a_j}\left(\sum_{i=0}^{n} \left(a_0\phi_0(x_i) + \cdots + a_m\phi_m(x_i) - f_i\right)^2\right) \\
&= 2\left(\sum_{i=0}^{n} \phi_j(x_i)\Big(a_0\phi_0(x_i) + \cdots + a_m\phi_m(x_i) - f_i\Big)\right) \\
&= 2\left(a_0\sum_{i=0}^{n}\phi_j(x_i)\phi_0(x_i) + \cdots + a_m\sum_{i=0}^{n}\phi_j(x_i)\phi_m(x_i) - \sum_{i=0}^{n}\phi_j(x_i)f_i\right) \\
&= 2\left(a_0\left(\phi_j,\phi_0\right) + \cdots + a_m\left(\phi_j,\phi_m\right) - \left(\phi_j,f\right)\right)
\end{aligned}$$

para $j = 0, 1, \ldots, n$. Assim, as condições

$$\frac{\partial J}{\partial a_j}(a_0, a_1, \ldots, a_m) = 0, \quad \forall j = 0, 1, \ldots m$$

implicam que

$$a_0\Big(\phi_j, \phi_0\Big) + a_1\Big(\phi_j, \phi_1\Big) + \cdots + a_m\Big(\phi_j, \phi_m\Big) = \Big(\phi_j, f\Big)$$

para $j = 0, 1, \ldots m$, de onde sai o sistema linear (5.23). □

Nota 5.39. A matriz do sistema (5.23) é simétrica, pelo que basta calcular a parte triangular inferior ou superior.

Nota 5.40. O caso $m = n$ (em que o número de pontos coincide com o número de funções base) corresponde a interpolação pelas funções de base ϕ_i considerando $i = 0, 1, \ldots, m$. Se $\phi_i(x) = x^i, \quad i = 0, 1, \ldots, n$, temos interpolação polinomial conforme tratado na secção 5.2.

Nota 5.41. Existe uma forma computacionalmente mais eficiente de calcular a matriz A e o vector b do segundo membro do sistema linear (5.23). Sendo C a matriz $(n+1) \times (m+1)$ dada por

$$C = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \ldots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_m(x_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \ldots & \phi_m(x_n) \end{bmatrix},$$

então

$$A = C^T C, \quad b = C^T F$$

em que F é o vector

$$F = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}.$$

Exercício 5.42. Um sinal de radio foi medido em vários instantes t, obtendo-se a seguinte tabela:

x_i	0	2	4	6
f_i	-0.0883	-0.766	0.954	-0.745

Determine a função da forma

$$\tilde{f}(x) = a_0 \sin x + a_1 \sin(2x)$$

que melhor aproxima os pontos dados.

Resposta.

Temos as funções de base

$$\phi_0(x) = \sin(x), \quad \phi_1(x) = \sin(2x),$$

e os pontos

$$x_0 = 0,\ x_1 = 2,\ x_2 = 4,\ x_3 = 6,$$

e os correspondentes valores de f

$$f_0 = -0.0883,\ f_1 = -0.766,\ f_2 = 0.954,\ f_3 = -0.745.$$

Assim a matriz C é dada por

$$C = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) \\ \phi_0(x_1) & \phi_1(x_1) \\ \phi_0(x_2) & \phi_1(x_2) \\ \phi_0(x_3) & \phi_1(x_3) \end{bmatrix} = \begin{bmatrix} 0. & 0. \\ 0.909297 & -0.756802 \\ -0.756802 & 0.989358 \\ -0.279415 & -0.536573 \end{bmatrix}$$

pelo que

$$A = C^T C = \begin{bmatrix} 1.47764 & -1.28698 \\ -1.28698 & 1.83949 \end{bmatrix}, \quad b = C^T F = \begin{bmatrix} -1.21004 \\ 1.92333 \end{bmatrix}.$$

Outra forma de obter o sistema é através da tabela:

x_i	$\Phi_0(x_i)$	$\Phi_1(x_i)$	$\Phi_0(x_i)^2$	$\Phi_0(x_i)\Phi_1(x_i)$	$\Phi_1(x_i)^2$	f_i	$f_i\Phi_0(x_i)$	$f_i\Phi_1(x_i)$
0	0	0	0	0	0	-0.0883	0	0
2	0.909297	-0.756802	0.826822	-688158	0.572750	-0.766	-0.696522	0.579710
4	-0.756802	0.989358	0.572750	-0.748748	0.978830	0.954	-0.721989	0.943848
6	-0.279415	-0.536573	0.078073	0.149927	0.287910	-0.745	0.208164	0.399747
$\sum$	-	-	1.47764	-1.28698	1.83949	-	-1.21004	1.92333

Resolvendo o sistema

$$A \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = b$$

obtemos

$$a_0 = 0.23491, a_1 = 1.20993$$

pelo que a aproximação da função é dada por

$$\tilde{f}(x) = 0.23491 \sin(x) + 1.20993 \sin(2x)$$

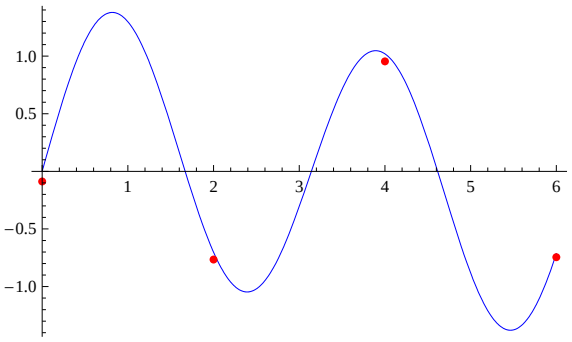


Figura 5.9: *Função aproximada por mínimos quadrados do exercício 5.42.*

Nota 5.43. Ao contrário da interpolação, quando aproximamos uma função por mínimos quadrados esta não passa necessariamente nos pontos dados. Isto vem do facto de o sistema ser sub-determinado no caso de se exigir que a função passasse nos pontos, uma vez que em geral o número de funções de base é inferior ao número de nós. No caso em que estes números coincidem temos interpolação com as funções de base consideradas, conforme já referido na nota 5.40.

Exercício 5.44. Alguns valores da função y foram medidos (com erro ligeiro) em vários pontos, conforme a tabela.

x_i	0	1	2	3
y_i	2.86	5.58	7.32	-0.171

.

Determine uma aproximação da função y, sabendo que esta é solução da equação diferencial de coeficientes constantes homogénea

$$y'' - 2y' + y = 0,$$

ou seja que a solução é da forma (ver [1, 3])

$$y(x) = C_1 e^x + C_2 x e^x,$$

para algum $C_1, C_2 \in \mathbb{R}$.

Resposta.

Para determinar as constantes, consideramos a aproximação por mínimos quadrados. Assim consideramos as funções de base

$$\phi_0(x) = e^x, \quad \phi_1(x) = x e^x,$$

os pontos

$$x_0 = 0,\ x_1 = 1,\ x_2 = 2,\ x_3 = 3$$

e os correspondentes valores de f

$$y_0 = 2.86,\ y_1 = 5.58,\ y_2 = 7.32,\ y_3 = -0.171.$$

Assim a matriz C é dada por

$$C = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) \\ \phi_0(x_1) & \phi_1(x_1) \\ \phi_0(x_2) & \phi_1(x_2) \\ \phi_0(x_3) & \phi_1(x_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ e & e \\ e^2 & 2e^2 \\ e^3 & 3e^3 \end{bmatrix}$$

pelo que

$$A = C^T C = \begin{bmatrix} 466.416 & 1326.87 \\ 1326.87 & 3856.64 \end{bmatrix}, \quad b = C^T F = \begin{bmatrix} 68.7 \\ 113. \end{bmatrix}.$$

Outra forma de obter o sistema é através da tabela:

x_i	$\Phi_0(x_i)$	$\Phi_1(x_i)$	$\Phi_0(x_i)^2$	$\Phi_0(x_i)\Phi_1(x_i)$	$\Phi_1(x_i)^2$	y_i	$y_i\Phi_0(x_i)$	$y_i\Phi_1(x_i)$
0	1	0	1	0	0	2.86	2.86	0
1	e	e	e^2	e^2	e^2	5.58	5.58 e	5.58 e
2	e^2	$2e^2$	e^4	$2e^4$	$4e^4$	7.32	7.32 e^2	14.64 e^2
3	e^3	$3e^3$	e^6	$3e^6$	$9e^6$	-0.171	-0.171 e^3	-0.513 e^3
$\sum$	-	-	466.416	1326.87	3856.64	-	68.7	113.

Resolvendo o sistema

$$A \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = b$$

obtemos

$$a_0 = 2.96, a_1 = -0.984$$

pelo que a aproximação da função é dada por

$$\tilde{y}(x) = 2.96e^x - 0.984e^x x.$$

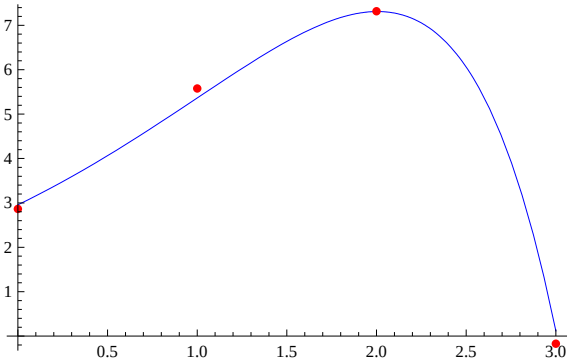


Figura 5.10: *Função aproximada por mínimos quadrados do exercício 5.44.*

Exercício 5.45. Alguns valores da função y foram medidos (com erro ligeiro) em vários pontos, conforme a tabela.

x_i	0	2	4
y_i	0.0931	7.06	54.6

.

Determine uma aproximação da função y, sabendo que esta é solução da equação homogénea

$$y'' - y = 0,$$

ou seja, é da forma $y(x) = C_1 e^x + C_2 e^{-x}$.

Resposta.

Temos a aproximação por mínimos quadrados

$$\tilde{y}(x) = 1.00e^x - 0.934e^{-x}.$$

Muitas vezes não sabemos qual o comportamento da função f e logo é difícil estabelecer um critério para escolher as funções de base. Nesse caso, geralmente aproximam-se os pontos dados por um polinómio. Assim as funções base são monómios, isto é,

$$\phi_j(x) = x^j, \quad j = 0, 1, \ldots, m.$$

Vamos de seguida estudar estes casos, começando com a aproximação por retas. Como veremos também, nesses casos particulares poderemos usar a função `polyfit` do Octave para fazer regressão polinomial.

5.4.1 Regressão Linear

Dada uma nuvem de pontos $(x_i, y_i)\, i = 0, 1, \ldots, n$, correspondentes às variáveis x e y, queremos determinar se existe alguma relação entre elas. Para isso, calcula-se a reta de regressão linear, no contexto da Estatística. No nosso contexto, a reta de regressão é nada mais nada menos que a reta que melhor aproxima os pontos $(x_i, y_i)\, i = 0, 1, \ldots, n$ no sentido dos mínimos quadrados. Assim, queremos determinar os coeficientes a_0, a_1 tal que a aproximação

$$\tilde{f}(x) = a_0 + a_1 x$$

aproxime os pontos dados. Neste caso temos as funções de base

$$\phi_0(x) = 1, \quad \phi_1(x) = x$$

pelo que a matriz A e o vector b do sistema (5.23) podem ser dados por

$$A = \begin{bmatrix} (\phi_0, \phi_0) & (\phi_0, \phi_1) \\ (\phi_1, \phi_0) & (\phi_1, \phi_1) \end{bmatrix} = \begin{bmatrix} n+1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix}$$

e

$$b = \begin{bmatrix} (\phi_0, f) \\ (\phi_1, f) \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} f_i \\ \sum_{i=0}^{n} f_i x_i \end{bmatrix}.$$

Assim, o sistema a resolver é dado por

$$\begin{bmatrix} n+1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} f_i \\ \sum_{i=0}^{n} f_i x_i \end{bmatrix}.$$

Exercício 5.46. Determine a reta que melhor aproxima os pontos da tabela no sentido dos mínimos quadrados.

x_i	0	1	2	3	4
f_i	0.966	1.11	1.50	1.90	1.78

Determine também o seu erro nos pontos dados no sentido dos mínimos quadrados.

Resposta.

Começamos por uma proposta de resolução manual. Para uma resolução em Octave, remetemos para o exercício 5.48 no final desta secção.

Construímos a tabela

	x_i	x_i^2	f_i	$x_i f_i$
	0	0	0.966	0
	1	1	1.11	1.11
	2	4	1.50	3.00
	3	9	1.90	5.71
	4	16	1.78	7.11
$\sum$	10	30	7.25	16.9

logo o sistema linear a resolver é

$$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 7.25 \\ 16.9 \end{bmatrix}$$

que tem a solução

$$a_0 = 0.966, \quad a_1 = 0.242$$

logo a reta de regressão linear é dada por

$$\tilde{f}(x) = 0.966 + 0.242x.$$

Para determinar o erro, temos de calcular

$$J(a_0, a_1) = \sum_{i=0}^{5} (\tilde{f}(x_i) - f_i)^2$$

logo com o auxílio da tabela

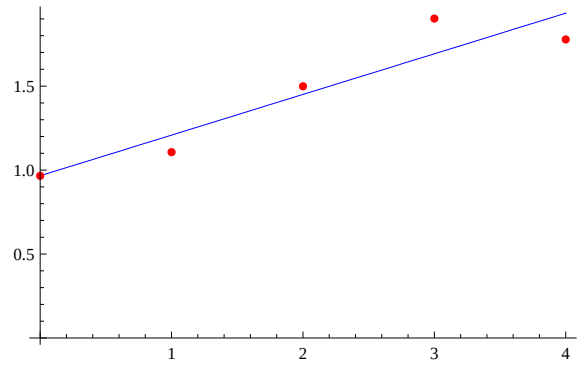


Figura 5.11: *Reta de regressão linear referente ao exercício 5.46.*

	x_i	f_i	$\tilde{f}(x_i)$	$\tilde{f}(x_i) - f_i$	$(\tilde{f}(x_i) - f_i)^2$
	0	0.966	0.966	0.000	0.000
	1	1.11	1.21	0.10	0.010
	2	1.50	1.45	-0.05	0.002
	3	1.90	1.69	-0.21	0.04
	4	1.78	1.93	0.16	0.02
$\sum$	10	7.25	-	-	0.08

obtemos o erro de 0.08 no sentido dos mínimos quadrados.

Exercício 5.47. Determine a reta de regressão linear, melhor adaptada aos pontos da tabela.

x_i	0.	0.5	1.	1.5
f_i	0.0139	0.283	0.582	0.968

Determine também o seu erro nos pontos dados no sentido dos mínimos quadrados.

Resposta.

A reta de regressão linear é

$$\tilde{f}(x) = 0.632515x - 0.0125367$$

e o seu erro no sentido dos mínimos quadrados é 0.00360384.

Exercício Octave 5.48.

A função `polyfit(lx,ly,n)` determina o polinómio de grau `n` que melhor se ajusta à lista de pontos de abcissas `lx` e ordenadas `ly` no sentido dos mínimos quadrados. Assim, determine as retas de regressão dos exercícios anteriores utilizando o comando `polyfit(lx,ly,1)`.

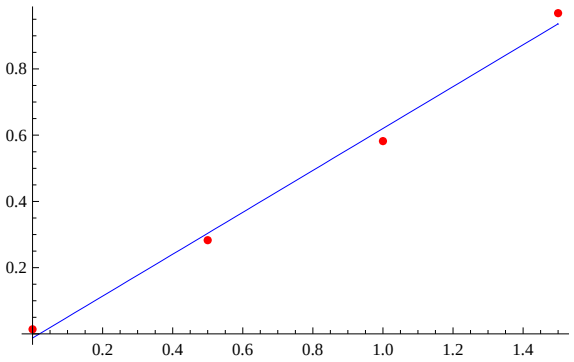


Figura 5.12: *Reta de regressão linear referente ao exercício 5.47.*

5.4.2 Regressão Exponencial

Podemos também utilizar a regressão linear para adaptar o gráfico de uma função da forma

$$f(x) = ae^{bx}$$

a um conjunto de $n+1$ pontos $\{(x_i, f_i) : \quad i = 0, 1, \ldots n\}$. Para isso basta proceder a uma mudança de variável. Assim, temos

$$\begin{aligned} f(x) = ae^{bx} \quad &\Leftrightarrow \quad \underbrace{\ln\Big(f(x)\Big)}_{g(x)} = \ln\left(ae^{bx}\right) \\ &\Leftrightarrow \quad g(x) = \ln a + \ln\left(e^{bx}\right) \\ &\Leftrightarrow \quad g(x) = \underbrace{\ln a}_{a_0} + \underbrace{b}_{a_1} x, \end{aligned}$$

ou seja, podemos procurar a função da forma

$$g(x) = a_0 + a_1 x$$

que melhor se adapta aos pontos $\left\{\big(x_i, g_i\big) : \quad i = 0, 1, \ldots n\right\}$, notando que $g_i = \ln(f_i)$ e que

$$a = e^{a_0}, \quad b = a_1.$$

Exercício 5.49. Determine a função da forma

$$f(x) = a\, e^{bx}$$

que melhor se adapta aos pontos tabelados.

x_i	0.1	0.3	0.7	1.3
f_i	1.43	1.11	0.616	0.230

Resolução. Considerando

$$g(x) = \ln(f(x))$$

temos a tabela

x_i	0.1	0.3	0.7	1.3
g_i	0.359	0.104	-0.485	-1.47

Aplicando regressão linear aplicada à tabela anterior, obtemos

$$a_0 = 0.547295, \quad a_1 = -1.53435,$$

e como

$$a = e^{a_0} = 1.72857, \quad b = a_1 = -1.53435,$$

obtemos a função

$$f(x) = 1.72857e^{-1.53435x}.$$

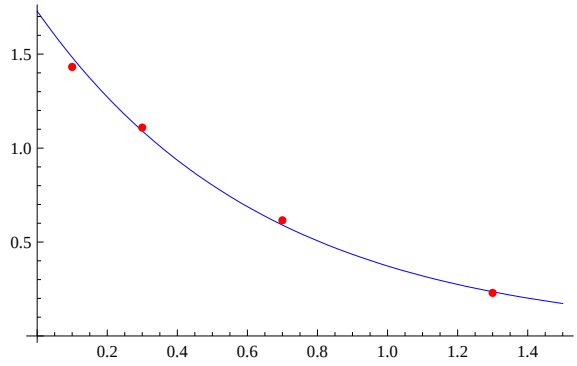


Figura 5.13: *Regressão exponencial referente ao exercício 5.49.*

Exercício 5.50. Determine a função da forma

$$f(x) = a\, e^{bx}$$

que melhor se adapta aos pontos tabelados.

x_i	1.1	2.3	3.7	4.3
f_i	1.56	2.40	5.83	9.04

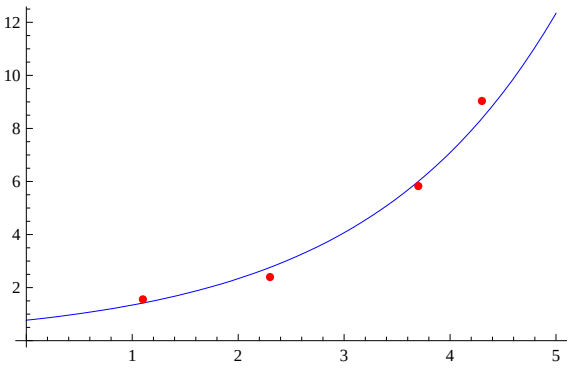


Figura 5.14: *Regressão exponencial referente ao exercício 5.50.*

Resposta.
$f(x) = 0.769952e^{0.554871x}$.

Exercício 5.51. Determine a função da forma

$$f(x) = a\, e^{bx}$$

que melhor se adapta aos pontos tabelados.

x_i	1.	1.2	1.4
f_i	7.44	10.7	16.7

Resposta.
$f(x) = 0.963895e^{2.02822x}$.

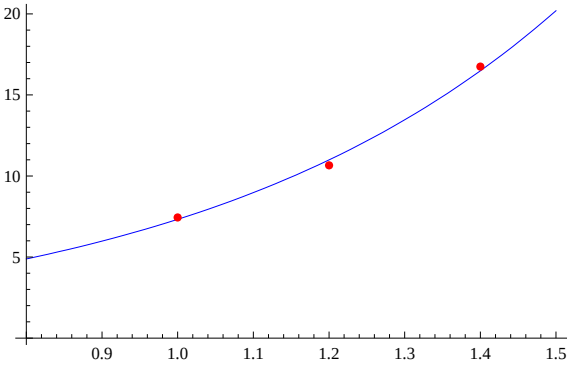


Figura 5.15: *Regressão exponencial referente ao exercício 5.51.*

5.4.3 Regressão Polinomial

Muitas vezes não temos informação sobre o tipo de comportamento da função. Nesse caso, o usual é fazer-se interpolação polinomial, isto é, considerar as funções de base

$$\phi_j(x) = x^j, \quad j = 0, 1, \ldots m,$$

em que o grau m é fixo. Assim, procuramos uma função aproximadora $\tilde{f}$ que seja um polinómio de grau m, ou seja,

$$\tilde{f}(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m.$$

A escolha do grau pode ser feita *a priori* pelo aspecto da nuvem de pontos ou *a posteriori* pela evolução do erro obtido e da suavidade da curva aproximada. À semelhança do obtido para regressão linear, a matriz A e o vector b do sistema linear (5.23) a resolver têm uma forma especial, nomeadamente

$$A = \begin{bmatrix} n+1 & \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \cdots & \sum_{i=0}^{n} x_i^m \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 & \cdots & \sum_{i=0}^{n} x_i^{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{n} x_i^m & \sum_{i=0}^{n} x_i^{m+1} & \sum_{i=0}^{n} x_i^{m+2} & \cdots & \sum_{i=0}^{n} x_i^{2m} \end{bmatrix}$$

e

$$b = \begin{bmatrix} \sum_{i=0}^{n} f_i \\ \sum_{i=0}^{n} f_i x_i \\ \vdots \\ \sum_{i=0}^{n} f_i x_i^m \end{bmatrix}.$$

Assim, o sistema a resolver é dado por

$$\begin{bmatrix} n+1 & \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \dots & \sum_{i=0}^{n} x_i^m \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 & \dots & \sum_{i=0}^{n} x_i^{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{n} x_i^m & \sum_{i=0}^{n} x_i^{m+1} & \sum_{i=0}^{n} x_i^{m+2} & \dots & \sum_{i=0}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} f_i \\ \sum_{i=0}^{n} f_i x_i \\ \vdots \\ \sum_{i=0}^{n} f_i x_i^m \end{bmatrix}. \quad (5.24)$$

Exercício 5.52. Considere os pontos tabelados e apresentados no gráfico da Figura 5.16.

x_i	1.1	1.4	1.7	2.	2.3
f_i	1.19	0.135	-0.0789	0.0202	0.328

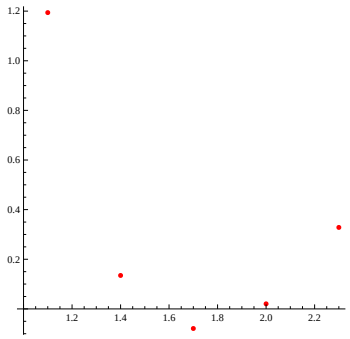


Figura 5.16: *Pontos tabelados referentes ao exercício 5.52.*

(a) Dada a distribuição dos pontos (x_i, f_i), considera melhor aproximar a função por um polinómio de grau 0, 1 ou 2?

(b) Determine a aproximação referente à alínea anterior, no sentido dos mínimos quadrados.

Resposta.

(a) Como um polinómio de grau zero representa uma reta horizontal, um polinómio de grau um representa uma reta não-vertical e um polinómio de grau 2 representa um parábola, dada a distribuição dos pontos o aconselhável é aproximar por um polinómio de grau 2.

(b) Temos então a seguinte tabela

	x_i	x_i^2	x_i^3	f_i	$f_i x_i$	$f_i x_i^2$
	1.1	1.21	1.331	1.19	1.31333	1.44466
	1.4	1.96	2.744	0.135	0.188715	0.2642
	1.7	2.89	4.913	-0.0789	-0.134108	-0.227984
	2.	4.	8.	0.0202	0.0404654	0.0809307
	2.3	5.29	12.167	0.328	0.755338	1.73728
$\sum$	8.5	15.35	29.155	1.60	2.16374	3.29908

logo de (5.24) obtemos o sistema linear

$$\begin{bmatrix} 5 & 8.5 & 15.35 \\ 8.5 & 15.35 & 29.155 \\ 15.35 & 29.155 & 57.6419 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1.59848 \\ 2.16374 \\ 3.29908 \end{bmatrix}$$

que tem solução

$$a_0 = 7.91993, \quad a_1 = -8.83842, \quad a_2 = 2.41859.$$

Assim o polinómio de grau 2 que melhor aproxima os pontos dados é

$$\tilde{f}(x) = 2.41859x^2 - 8.83842x + 7.91993.$$

Exercício 5.53. Considere a seguinte tabela:

x_i	1.03	1.34	1.87
f_i	1.39	1.53	2.66

.

(a) Determine o polinómio de grau 2 que melhor aproxima os pontos no sentido dos mínimos quadrados.

(b) Determine o polinómio interpolador correspondente aos pontos tabelados.

(c) Determine o erro no sentido dos mínimos quadrados cometido na aproximação da alínea $a)$.

Resposta. (a) $\tilde{f}(x) = 2.00764x^2 - 4.31496x + 3.70645.$

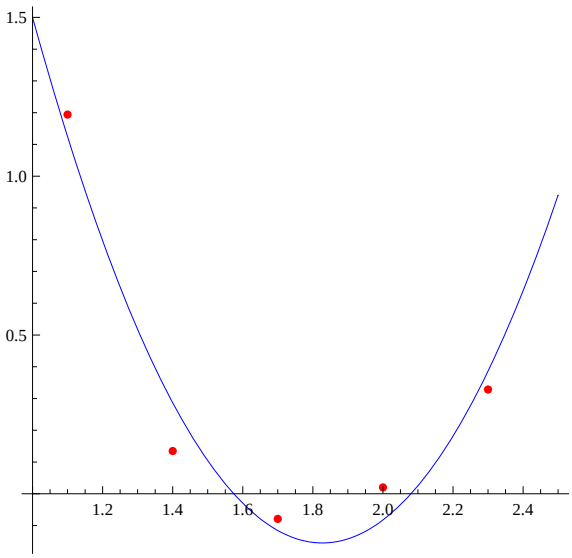


Figura 5.17: *Regressão polinomial de grau 2 referente ao exercício 5.52.*

(b) Como 3 nós determinam um polinómio interpolador de grau menor ou igual a 2 e o polinómio interpolador satisfaz $p_2(x_i) = f_i$, a função obtida por aproximação por mínimos quadrados coincide com o polinómio interpolador de grau 2, isto é,

$$p_2(x) = \tilde{f}(x) = 2.00764x^2 - 4.31496x + 3.70645.$$

(c) Como a função $\tilde{f}$ coincide com o polinómio interpolador de grau 2, temos em particular que

$$\tilde{f}(x_i) = f_i,$$

logo pela definição de erro no sentido dos mínimos quadrados (5.22), o erro é nulo.

Exercício 5.54. Considere a tabela seguinte

x_i	-0.5	0	0.5	1
f_i	0.513	1.45	2.58	3.59

.

(a) Determine o polinómio de grau 2 que melhor se adequa aos pontos tabelados.

(b) Indique se o grau do polinómio escolhido é apropriado.

Resposta. (a) $\tilde{f}(x) = 0.064205x^2 + 2.03947x + 1.50019.$

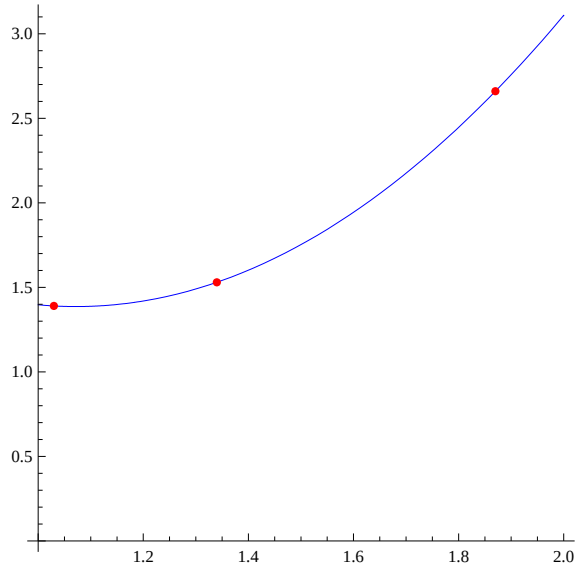


Figura 5.18: *Regressão polinomial de grau 2 referente ao exercício 5.52. Como se verifica, este coincide com interpolação polinomial, pois a aproximação $\tilde{f}$ passa nos pontos dados.*

(b) Como o coeficiente do termo de segunda ordem é próximo de zero, bastaria talvez fazer regressão linear. De facto, o gráfico de $\tilde{f}$ é praticamente uma reta (ver Figura 5.19).

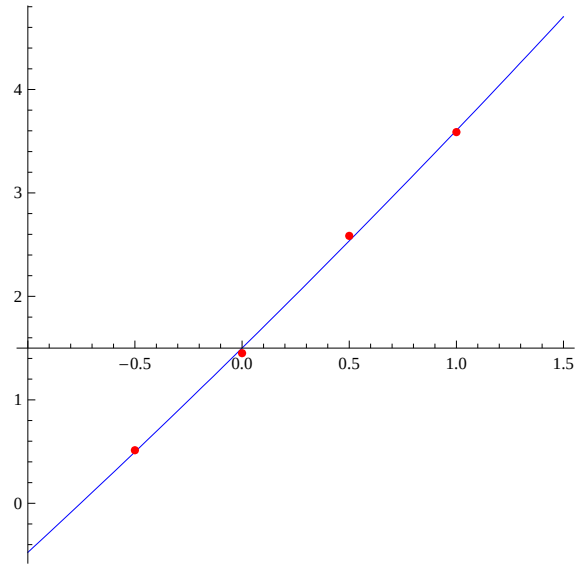


Figura 5.19: *Regressão polinomial de grau 2 referente ao exercício 5.54.*

Exercício Octave 5.55.

A função `polyfit(lx,ly,n)` determina o polinómio de grau `n` que melhor se ajusta à lista de pontos de abcissas `lx` e ordenadas `ly` no sentido dos mínimos quadrados. Assim, à semelhança do exercício 5.48, determine as respostas dos exercícios anteriores utilizando o comando `polyfit`.

Capítulo 6

Métodos Iterativos para Equações Não-Lineares

Neste capítulo vamos estudar formas de resolução numérica de equações não lineares. Note-se que este problema é equivalente ao de encontrar raízes de uma função, uma vez que

$$g(x) = b \quad \Leftrightarrow f(x) = 0, \quad \text{com } f(x) = g(x) - b$$

Muitas vezes não é possível obter um processo de resolução direta que encontre a solução exata. Outras vezes esse processo existe, mas é computacionalmente muito moroso. Desta forma é necessário procurar formas alternativas de encontrar soluções aproximadas do problema em questão. Muitas das vezes isto passa por aplicar um método iterativo, que definimos de seguida.

Definição 6.1 (Método iterativo).

Seja $x^{(0)}$ a iterada ou aproximação inicial e $x^{(k)}$ a denominada k-ésima iterada.

Chama-se **método iterativo** a um procedimento sucessivo (chamado iteração) que a partir de $x^{(k)}$ gera uma nova aproximação $x^{(k+1)}$ da solução x do problema.

Caso a sucessão de iteradas $x^{(k)}$ convirja para a solução x, isto é,

$$\lim_{k\to\infty} x^{(k)} = x,$$

o método diz-se **convergente** . Caso contrário diz-se **divergente**.

Nota 6.2. Note-se que um método é convergente para a solução x do problema se o erro absoluto das iteradas $x^{(k)}$ sucessivas tender para zero, isto é, caso

$$||x^{(k)} - x|| \to 0, \quad k \to \infty.$$

Antes de apresentarmos alguns métodos iterativos para resolver equações não-lineares, vamos estabelecer as bases para esta secção. Começamos por recordar teoremas de unicidade e existência de raíz de f no intervalo $[a, b]$. As suas demonstrações podem ser encontradas em qualquer livro de introdução à análise real, como por exemplo [7].

Teorema 6.3 (Bolzano).
Seja $f : [a, b] \to \mathbb{R}$ contínua. Então se $f(a).f(b) \leq 0$ existe (pelo menos) uma raíz de f em $[a, b]$.

Teorema 6.4 (Lagrange).
Seja $f : [a, b] \to \mathbb{R}$ diferenciável. Então existe um $\xi \in [a, b]$ tal que

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

Corolário 6.5 (Rolle).
Seja $f : [a, b] \to \mathbb{R}$ diferenciável. Se $f(a) = f(b)$ então existe um $\xi \in [a, b]$ tal que $f'(\xi) = 0$.

O contra-recíproco do corolário de Rolle dá-nos o resultado de unicidade de raiz.

Teorema 6.6.
Seja $f : [a, b] \to \mathbb{R}$ diferenciável. Se $f'(x)$ não se anula $[a, b]$ então existe no máximo uma raíz de f em $[a, b]$.

Podemos então formular o seguinte resultado de existência e unicidade de raiz.

Corolário 6.7 (Existência e unicidade de raíz).
Seja $f : [a, b] \to \mathbb{R}$ diferenciável. Então se

- *$f'(x)$ não muda de sinal em $[a, b]$*
- *$f(a).f(b) \leq 0$*

a função f tem uma única raíz em $[a, b]$.

Pelo Teorema de Lagrange, temos também a estimativa seguinte para o erro absoluto na aproximação de raizes.

Corolário 6.8 (Estimativa do erro da aproximação de raízes).

Seja $f : [a, b] \to \mathbb{R}$ diferenciável e seja $z \in [a, b]$ uma raiz f. Para uma aproximação $\tilde{z}$ de z temos

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

A estimativa anterior é geral e não depende do método de aproximação de raízes.

Para completar os teoremas base para este capítulo, relembramos também o teorema de Weierstrass.

Teorema 6.9 (Weierstrass).

Seja $f : [a, b] \to \mathbb{R}$ uma função contínua no intervalo fechado $[a, b]$. Então f tem um máximo e um mínimo em $[a, b]$.

O método iterativo mais simples para encontrar raizes de uma função real é o método da bisseção, ao qual dedicamos a seguinte secção.

6.1 Método da bisseção

Dada uma função $f : \mathbb{R} \to \mathbb{R}$, o método da bisseção procura encontrar uma raíz de f num dado intervalo $[a, b]$, isto é, procura a solução de

$$f(x) = 0, \quad x \in [a, b].$$

Começamos então por expor o método da bisseção.

Teorema 6.10 (Método da Bisseção).

*Seja $f : [a, b] \to \mathbb{R}$ contínua, tal que $f(a).f(b) < 0$. Chama-se **método da bisseção** ao método iterativo com iterada inicial*

$$a_0 = a, \quad b_0 = b, \quad c_0 = \frac{a_0 + b_0}{2},$$

com iterações dadas por

$$a_{n+1} = \begin{cases} a_n, & \text{se } f(a_n).f(c_n) < 0 \\ c_n, & \text{caso contrário} \end{cases}, \quad b_{n+1} = \begin{cases} c_n, & \text{se } f(a_n).f(c_n) < 0 \\ b_n, & \text{caso contrário} \end{cases},$$

$$c_{n+1} = \frac{a_{n+1} + b_{n+1}}{2}, \tag{6.1}$$

para $n \in \mathbb{N}$. Então existe uma raiz $z \in [a,b]$ de f para a qual temos a estimativa de erro

$$|e_n| \leq \frac{b-a}{2^{n+1}}, \tag{6.2}$$

em que $e_n = z - c_n$ é o erro da iterada n.

O método da bisseção é talvez o método mais simples para determinar soluções de equações não lineares. A ideia é dividir o intervalo inicial sucessivamente ao meio e utilizar o teorema de Bolzano 6.3 para decidir em qual dos dois novos subintervalos está a raiz. No entanto tem algumas desvantangens. Em primeiro lugar, só pode ser aplicado a funções reais e a generalização para operadores mais complicados não é trivial. Em segundo lugar, é pouco eficiente. Veremos nas secções seguintes métodos mais eficientes para o fazer. Por outro lado, tem também algumas vantagens. Por um lado não exige que a função f seja diferenciável, bastando ser contínua. Depois pode ser aplicado mesmo sem garantia de unicidade de raíz, sendo que nesse caso convergirá para apenas uma das raízes. Mais ainda, a estimativa de erro é independente da função f. Isto mostra que a performance do método da bisseção não tem grande dependência em relação à função em causa, sendo portanto um método geral para resolução de equações não-lineares reais.

Exercício Octave 6.11. Crie uma função `bissec(expf,a,b,n)` que dada a expressão de $f(x)$ calcule a aproximação de ordem n pelo método da bisseção para a raíz de f em $[a,b]$.

Resposta.

```
function c = bissec(expf, a, b, niter)
% Método da bisseção
% Input:  expf - expressão de f
%   [a,b] - intervalo
%   niter - numero de iteracoes
% Output:  c - aproximacao da raiz
f = inline(expf);
if f(a)*f(b) > 0
  disp('f(a)*f(b)>0!  Nao se pode aplicar o metodo!')
elseif f(a)*f(b) == 0
  disp('f(a)*f(b)=0!  Ou a ou b é raiz de f!')
else
  for iter = 1:  niter+1
    c = (a+b)/2;
```

```
    if f(a)*f(c) <0
      b=c;
    else
      a=c;
    end
  end
end
return
```

Exercício Octave 6.12. Considere a equação $e^{-x} = 0.5$.

(a) Indique se está nas condições para aplicar o método da bisseção para resolver esta equação considerando o intervalo [0,1].

(b) Mostre que existe uma única solução da equação e que está nesse intervalo.

(c) Utilize o algoritmo do exercício 6.11 para determinar uma aproximação da solução da equação com 20 iterações do método da bisseção.

(d) Calcule um majorante para o erro cometido na alínea anterior.

(e) Para garantir que se obtem um erro absoluto inferior a 10^{-10}, quantas iterações seriam necessárias.

Resolução.

(a) Resolver a equação dada equivale a encontrar as raízes de $f(x) = e^{-x} - 0.5$. Como f é contínua e $f(0).f(1) = -0,066 < 0$ estamos nas condições de aplicação do método da bisseção.

(b) Como $f'(x) = -e^{-x} < 0$ para qualquer $x \in \mathbb{R}$, pelo Teorema de Rolle existe uma única raíz de f em $\mathbb{R}$. Em conjunto com a alínea anterior, estamos nas condições do corolário 6.7, logo a raiz de f está no intervalo $[0, 1]$.

(c) Aplicando `bissec('exp(-x)-0.5', 0, 1, 20)`, obtemos a aproximação $c_{20} = 0.69315$.

(d) De (6.2) temos a estimativa $e_{20} \leq 4.7684 \times 10^{-7}$

(e) De (6.2), temos

$$e_n \leq \frac{b-a}{2^{n+1}} \leq 10^{-10} \Leftrightarrow 2^{n+1} \geq 10^{10} \Leftrightarrow n \geq \log_2(10^{10}) - 1 \approx 32.219$$

logo seriam necessárias $n = 33$ iterações.

O método da bisseção é bastante intuitivo e fácil de utilizar. No entanto tem desvantagens, algumas das quais já descritas. Outra particularidade pouco abonatória é que o método não garante que a aproximação seguinte seja menor que a anterior. Em particular, para a função $f(x) = x - 0.499$ que tem uma única raiz $z = 0.499$, a aplicação do método da bisseção no intervalo [0,1] dá-nos as aproximações consecutivas

$$c_0 = 0.5,\ c_1 = 0.25,\ c_2 = 0.375,\ c_3 = 0.4375, \ldots, c_7 = 0.49609,\ c_8 = 0.49802.$$

Em particular o erros da aproximação c_0 é menor do que todos os erros até c_8, ilustrando o que acabámos de dizer.

Nas secções seguintes vamos apresentar outros iterativos métodos para resoluções de equações não lineares que garantem melhor performance no que se refere à rapidez de convergência.

6.2 Método do ponto fixo

Iniciámos este capítulo com o problema de encontrar raízes de funções não lineares. Nesta secção vamos estudar o problema de encontrar pontos fixos de equações não lineares. Para iniciar a análise vamos considerar pontos fixos de uma função real $g : I \subset \mathbb{R} \to \mathbb{R}$.

Começamos com a definição de ponto fixo.

Definição 6.13 (Ponto Fixo).
Diz-se que $x \in I \subset \mathbb{R}$ é **ponto fixo** de $g : I \to \mathbb{R}$ se $g(x) = x$.

Note-se que de certa forma o problema de encontrar pontos fixos de uma função g é equivalente ao problema de encontrar raízes da função f com expressão $f(x) = g(x) - x$.

Para continuar com a nossa análise é útil a definição de contração.

Teorema 6.14 (Contração).
*Seja $I \subset \mathbb{R}$. Diz-se que a função $g : I \to \mathbb{R}$ é uma **contração** se existir uma constante $0 \leq L < 1$ tal que*

$$|g(x) - g(y)| \leq L|x - y|$$

para qualquer $x, y \in I$. A L chama-se constante de contração.

Nesta fase convém fazer duas notas. A primeira é que qualquer contração é uma função contínua, uma vez que $x_n \to x$, $n \to \infty$ que

$$|g(x) - g(x_n)| \leq L|x - x_n| \to 0, \quad n \to \infty.$$

A segunda é que uma contração é um caso particular de uma função Lipschitz. diz-se que uma função é Lipschitz (ou Lipschitz contínua) se existe $L \in \mathbb{R}$ tal que

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

Assim sendo, uma contração é uma função Lipschitz com constante de Lipschitz $L < 1$.

Temos então o seguinte teorema, que é o resultado principal desta secção. Além de mostrar a existência e unicidade de ponto fixo sob determinadas condições, a sua demonstração é construtiva, indicando um método de aproximação do ponto fixo.

Teorema 6.15 (Ponto Fixo em $\mathbb{R}$).
Seja $I \subset \mathbb{R}$ um conjunto completo.
Seja $g : I \to I$ uma contração com constante de contração $L < 1$.
Então existe um único ponto fixo x de g em I.

Demonstração. Começamos por mostrar a unicidade. Por absurdo, supomos que existem dois pontos fixos distintos x_1, x_2 tal que $g(x_i) = x_i$, $i = 1, 2$. Então, como g é contração temos

$$|x_1 - x_2| = |g(x_1) - g(x_2)| \leq L|x_1 - x_2| < |x_1 - x_2|,$$

o que é absurdo, pelo que mostrámos unicidade.

Para mostrar a existência de ponto fixo, começamos por considerar a sucessão $(x_n)_{n\in\mathbb{N}}$ definida por

$$x_{n+1} = g(x_n), \quad n \in \mathbb{N}_0$$

com $x_0 \in I$ arbitrário. Vamos mostrar que a sucessão é convergente e que o seu limite é ponto fixo de g. Temos

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq L|x_n - x_{n-1}| < |,$$

para qualquer $n \in \mathbb{N}_0$, logo por indução temos

$$|x_{n+1} - x_n| \leq L^n|x_1 - x_0|.$$

Além disso, por um argumento semelhante, temos que para $m > n$

$$|x_m - x_n| \leq L^n |x_{m-n} - x_0|,$$

e pela desigualdade triangular

$$|x_{m-n} - x_0| \leq |x_1 - x_0| + |x_2 - x_1| + \cdots + |x_{m-n-1} - x_{m-n}| \leq \sum_{k=0}^{m-n-1} L^k |x_1 - x_0|.$$

Tomando a soma de uma série geométrica com $L < 1$, temos

$$\sum_{k=0}^{n-m-1} L^k \leq \sum_{k=0}^{\infty} L^k \leq \frac{1}{1-L}$$

logo obtemos

$$|x_m - x_n| \leq \frac{L^n}{1-L} |x_1 - x_0|. \tag{6.3}$$

Como $L^n \to 0$ para $n \to \infty$, temos que a sucessão $(x_n)_{n\in\mathbb{N}}$ é de Cauchy, logo tem limite $x \in I$, uma vez que I é completo. Assim, a prova fica terminada notando que

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

uma vez que qualquer contração é um operador contínuo. □

O Teorema anterior é construtivo, no sentido em que a sua demonstração indica um método iterativo que permite aproximar o ponto fixo $x = g(x)$ de g com precisão arbitrária, quando as condições do teorema são verificadas. A este método iterativo, chama-se método do ponto fixo.

Teorema 6.16 (Método do Ponto Fixo em $\mathbb{R}$).

Nas condições do teorema 6.15, o ***método do ponto fixo*** *definido por uma iterada inicial $x_0 \in I$ e pela iteração*

$$x_{n+1} = g(x_n)$$

para $n \in \mathbb{N}_0$ converge para o único ponto fixo de g em I. Mais ainda, temos as estimativas de erro a priori

$$|e_n| \leq \frac{L^n}{1-L} |x_1 - x_0| \tag{6.4}$$

e a posteriori

$$|e_n| \leq \frac{L}{1-L} |x_n - x_{n-1}|, \tag{6.5}$$

em que $e_n = x - x_n$ é o erro da iterada n.

Demonstração. A demonstração está contida na demonstração do teorema 6.15 à excepção das estimativas de erro. A estimativa (6.4) sai diretamente de (6.3) fazendo nessa desigualdade o limite $m \to \infty$. A estimativa (6.5) sai de (6.4) considerando a iterada inicial x_{n-1}. □

Convém nesta altura fazer algumas considerações. Em relação às estimativas, note-se que a estimativa *a posteriori* (6.5) é sempre tão boa ou melhor que a estimativa *a priori* (6.4), uma vez que L é um majorante para o ganho em termos de erro de uma iteração. Assim, sempre que possível deve se usar a estimativa *a posteriori* (6.5) em detrimento da estimativa *a priori* (6.4). Por outro lado, a estimativa (6.5) necessita que sejam calculadas iterações do método e tal como a designação indica não pode ser calculada *a priori*, antes de se iniciar as iterações do método.

Por outro lado, verificamos que as hipóteses base para o teorema do Ponto Fixo são três: $I \in \mathbb{R}$ é completo, g é contração e $g(I) \subset I$ (ou seja, a imagem de I por g está contida em I). Quanto à primeira hipótese, sabemos que que o conjunto $\mathbb{R}$ é completo [7]. De igual forma, qualquer intervalo limitado e fechado é completo, assim como qualquer intervalo da forma $[a, \infty[$ ou $[\infty, a[$ para $a \in \mathbb{R}$. Desta forma a hipótese I completo em 6.29 é satisfeita em qualquer destas situações. Quanto à segunda condição, o teorema de Lagrange 6.4 garante que se g é diferenciável e

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

então g é automaticamente uma contração. Nessa perspectiva, esta é uma condição suficiente para que g seja uma contração. Assim sendo, temos o seguinte corolário, cujas hipóteses, embora não sejam tão abrangentes como as do Teorema do Ponto Fixo 6.15, são geralmente de mais fácil verificação.

Corolário 6.17 (Método do Ponto Fixo em $\mathbb{R}$).
Seja $I = [a, b]$ um intervalo limitado e fechado ou $I = \mathbb{R}$. Seja $g : I \to \mathbb{R}$ uma função diferenciável. Então se

(a) $g(I) \subset I$;

(b) $L = \max_{x \in I} |g'(x)| < 1$

o método do ponto fixo $x_{n+1} = g(x_n)$ converge para o único ponto fixo x de g em I, i.e.,

$$g(x) = x.$$

Mais ainda, são válidas as estimativas a priori (6.4) *e* a posteriori (6.5).

Exercício Octave 6.18. Escreva uma função em Octave

`pontofixoreal1(expg,x0,niter,L)`

que calcule `niter` iterações do método do ponto fixo para a função com expressão `expg` a partir da iterada inicial `x0`. Caso `L` seja introduzido como um valor $0 < L < 1$ para a constante de contração, a função deve indicar a estimativa *a priori* e *a posteriori* da iterada final.

Resposta.

Temos a proposta de algoritmo:

```
function [lx,Epri,Epos]=pontofixoreal1(expg, x0, niter, L)
% Calcula niter iteracoes do ponto fixo.
% INPUT: expg - expressão da função g
% x0 - iterada inicial
% niter - numero de iteracoes
% L - constante de contracao
% OUTPUT: lx - lista da iteradas
% Epri - estimativa a priori da iterada final
% Epos - estimativa a posteriori da iterada final
g = inline(expg);
lx = zeros(1,niter+1);
lx(1) = x0;
for iter = 1:  niter
  lx(iter+1) = g(lx(iter));
end
disp(['Iterada ',num2str(niter),' = ', num2str(lx(end))]);
if L <1 && L >0
  Epri = L^(niter)/(1-L)*abs(lx(2)-lx(1));
  Epos = L/(1-L)*abs(lx(end)-lx(end-1));
  disp(['Estimativa a priori:  |x-xn| <= ',num2str(Epri)]);
  disp(['Estimativa a posteriori:  |x-xn| <= ',num2str(Epos)]);
end
return
```

Exercício Octave 6.19. Escreva uma função em Octave

`pontofixoreal2(expg, x0, L, tol)`

que calcule iterações do método do ponto fixo para a função com expressão `expg` a partir da iterada inicial `x0`. O método deve parar quando a estimativa *a posteriori* garantir um erro de aproximação menor que a tolerância `tol`. A função deve indicar a estimativa *a priori* e *a posteriori* da iterada final.

Resposta.

No ficheiro `pontofixoreal2.m` escrevemos:

```
function [lx, Epri, Epos] = pontofixoreal2(expg, x0, L, tol)
% Calcula iteracoes do método do ponto fixo até erro < tol
% INPUT: expg - expressão da função g
% x0 - iterada inicial
% L - constante de contracao
% tol - tolerância
% OUTPUT: lx - lista da iteradas
% Epri - estimativa a priori da iterada final
% Epos - estimativa a posteriori da iterada final
g = inline(expg);
if tol <= 0
  disp('Tolerancia tol tem de ser positiva!');
elseif L <1 && L >0
  lx = [x0];
  iiter = 1;
  while iiter==1 || Epos>tol
    lx(iiter+1) = g(lx(end));
    iiter = iiter+1;
    Epos = L/(1-L)*abs(lx(end)-lx(end-1));
  end
  niter = iiter;
  Epri = L^(niter)/(1-L)*abs(lx(2)-lx(1));
  disp(['Estimativa a priori:  |x-xn| <= ',num2str(Epri)]);
  disp(['Estimativa a posteriori:  |x-xn| <= ',num2str(Epos)]);
else
  disp('A constante L tem se satisfazer 0 < L <1!')
end
return
```

Exercício 6.20. Considere $I = [0, 1]$ e função de expressão

$$f(x) = \sin(x + 1).$$

(a) Mostre que f tem um único ponto fixo em I.

(b) Considerando a iterada inicial $x_0 = 0$, determine uma estimativa *a priori* para o número de iterações necessárias para obter um erro inferior a 10^{-2}.

(c) Considerando a mesma iterada inicial, calcule o número de iterações necessárias de modo a garantir um erro inferior a 10^{-2} usando uma estimativa *a posteriori*.

Resposta.

(a) Temos I intervalo limitado e fechado, logo completo. Mais ainda,

$$f(I) = \sin([0,1]+1) = \sin([1,2]) \subset \sin([0,\pi]) = [0,1] = I.$$

Por fim, temos

$$L = \max_{x\in I} |g'(x)| = \max_{x\in[0,1]} |\cos(x+1)| = \max_{x\in[1,2]} |\cos(x)|.$$

Como a função coseno é decrescente em $[0,\pi] \supset [1,2]$, temos que

$$L = \max_{x\in[1,2]} |\cos(x)| = \max\{|\cos(1)|, |\cos(2)|\} = 0.54030 < 1,$$

logo f é uma contração e está nas condições do teorema do Ponto Fixo. Assim, f tem um único ponto fixo em I.

(b) Temos $x_1 = f(0) = 0.84147$. Assim pela estimativa *a priori* (6.4) temos

$$\begin{aligned} |e_n| \leq \frac{L^n}{1-L}|x_1 - x_0| < 10^{-2} &\Leftrightarrow L^n < \frac{(1-L)10^{-2}}{|x_1 - x_0|} \\ &\Leftrightarrow n \ln L < \ln\left(\frac{(1-L)10^{-2}}{|x_1 - x_0|}\right) \\ &\Leftrightarrow n > \frac{\ln\left(\frac{(1-L)10^{-2}}{|x_1 - x_0|}\right)}{\ln L} \\ &\Leftrightarrow n > 8.4625 \end{aligned}$$

logo são necessárias $n = 9$ iterações por uma análise *a priori*.

(c) Utilizando a estimativa *a posteriori* (6.5) temos

$$\begin{aligned}
x_2 = f(x_1) = 0.96359 \quad &\Rightarrow \quad |e_2| \leq \frac{L}{1-L}|x_2 - x_1| = 0.14353 \\
x_3 = f(x_2) = 0.92384 \quad &\Rightarrow \quad |e_3| \leq \frac{L}{1-L}|x_3 - x_2| = 0.046717 \\
x_4 = f(x_3) = 0.93832 \quad &\Rightarrow \quad |e_4| \leq \frac{L}{1-L}|x_4 - x_3| = 0.01702 \\
x_5 = f(x_4) = 0.93322 \quad &\Rightarrow \quad |e_5| \leq \frac{L}{1-L}|x_5 - x_4| = 0.0060001
\end{aligned}$$

logo bastam 5 iterações para se ter o majorante de erro desejado.

Por outro lado, a partir da derivada de g podemos obter uma condição de divergência do método do ponto fixo.

Teorema 6.21 (Divergência do método do ponto fixo).
Seja $I = [a, b]$ um intervalo limitado e fechado ou $I = \mathbb{R}$. Seja $g : I \to \mathbb{R}$ uma função diferenciável. Então se para o ponto fixo x de g se tem

$$|g'(x)| > 1 \tag{6.6}$$

o método do ponto fixo é localmente divergente (a não ser que uma da iteradas seja acidentalmente $x_n = x$.)

Demonstração. Se $|g'(x)| > 1$ existe um $\varepsilon > 0$ tal que para $x_n \in [x - \varepsilon, x + \varepsilon]$ e $x_n \neq x$, temos pelo Teorema de Lagrange 6.4 que existe um $xi \in [x_n; x]$ tal que

$$x_{n+1} - x = g(x_n) - g(z) = g'(\xi)(x_n - z)$$

logo passando ao módulo obtemos que

$$|x_{n+1} - x| = |g'(\xi)||x_n - z| > |x_n - z|$$

ou seja o erro aumenta com a iteradas em vez de diminuir. □

O teorema anterior mostra que se $|g'(x)| > 1$ então g não pode ser uma contração em torno de do ponto fixo x. Isto é ilustrado no exemplo seguinte.

Exercício 6.22. Considere $I = [-1, 0]$ e função de expressão

$$f(x) = x^4 - 1.$$

(a) Mostre que f não é uma contração em I.

(b) Mostre que f tem um único ponto fixo em $[-1, 0]$.

(c) Calcule 10 iterações do método do ponto fixo para a função anterior, considerando $x_0 = -0.5$. Que conclui?

Resposta.

(a) Temos $f'(x) = 4x^3$, logo parecem existir problemas em termos de contração (derivada maior que 1) para $x \approx -1$. Assim, escolhendo $x = -1, y = -0.9$, temos

$$|f(x) - f(y)| = |0 - (-0.34)| = 0.34 > 3|x - y|$$

logo não existe $L < 1$ tal que a condição de contração seja satisfeita para qualquer $x, y \in I$.

(b) Tomamos a função auxiliar $h(x) = f(x) - x$. Assim, f tem um único ponto fixo no intervalo se e só se h tem uma única raíz no intervalo. Como temos $h(-1) = 1$ e $h(0) = -1$, pelo teorema de Bolzano 6.3 temos que f tem pelo menos um ponto fixo no intervalo. Como

$$h'(x) = 0 \Leftrightarrow 4x^3 - 1 = 0 \Leftrightarrow x = \sqrt[3]{\frac{1}{4}},$$

ou seja, $h'(x) \neq 0$ no intervalo, concluímos a prova pelo Corolário 6.7.

(c) Como a função f verifica $|f'(x)| > 1$ para $-1 \leq x \leq -\sqrt[3]{1/4} \approx -0.62996$ e $f(-\sqrt[3]{1/4}) < 0$, concluímos que $z \in [-1, -\sqrt[3]{1/4}]$. Assim a condição de divergência (6.6) é satisfeita. Como ilustração, temos

$$\begin{aligned} x_1 &= -0.93750; \quad x_2 = -0.22752; \quad x_3 = -0.99732; \\ x_4 &= -0.01068; \quad x_5 = -1.0000; \quad x_6 = 0.0000; \\ x_7 &= -1.0000; \quad x_8 = 0.0000; \quad x_9 = -1.0000; \\ x_{10} &= 0.0000. \end{aligned}$$

ou seja, o método é divergente, oscilando com período dois entre os valores 0 e -1.

Nesta altura faz sentido comparar a performance do método da bisseção com o método do ponto fixo. Consideramos o problema do exercício 6.20 de encontrar a raiz de $f(x) = \sin(x + 1) - x$ para o método da bisseção no intervalo [0,1] e o problema de encontrar o ponto fixo de $g(x) = \sin(x + 1)$ para o método do ponto

fixo, que têm obviamente a mesma solução. Sabemos que a solução aproximada a 15 casas decimais é

$$x = 0.934563210752024.$$

Temos então a tabela 6.1, que compara a evolução das iterações e erros em cada caso.

n	Met. Bisseção		Met. Ponto Fixo	
	c_n	Erro	x_n	Erro
0	0.5000000000	0.434563	0.5000000000	0.434563
1	0.7500000000	0.184563	0.9974949866	0.0629318
2	0.8750000000	0.0595632	0.9103370262	0.0242262
4	0.9062500000	0.0283132	0.9315617495	0.00300146
8	0.9355468750	0.000983664	0.9345152492	4.79616×10^{-5}
16	0.9345626831	5.27647×10^{-7}	0.9345631984	1.23167×10^{-8}
32	0.9345632108	6.40983×10^{-11}	0.9345632108	8.88178×10^{-16}

Tabela 6.1: *Tabela de iterações do método da bisseção e do ponto fixo para determinar a solução de* $x = \sin(x + 1)$.

Reparamos portanto que a rapidez de convergência difere entre os dois métodos, sendo que o método do ponto fixo parece ser mais rápido que o da bisseção. Uma pergunta legítima nesta altura é saber se o método do ponto fixo é sempre mais rápido que o da bisseção ou se acontece apenas para este exemplo. Para responder a esta pergunta precisamos da definição de ordem de convergência. A ordem de convergência indica a relação entre os erros de duas iteradas consecutivas. Nomeadamente indica com que potência decresce o erro em duas iteradas consecutivas.

Definição 6.23 (Ordem de convergência).

Um método iterativo diz-se com **ordem de convergência** $p \geq 1$ se existir uma constante C tal que a sucessão $(x_n)_{\mathbb{N}}$ das iteradas convergente para a solução x satisfaz

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C.$$

em que $e_n = x - x_n$ é o erro da iterada x_n. Isto é equivalente a dizer que existe um $C > 0$ e um $N \in \mathbb{N}$ tal que para qualquer $n > N$ se tem

$$|e_{n+1}| \leq C|e_n|^p.$$

Um método que converge com ordem $p = 1$ diz-se que tem **convergência linear**. Um método que converge com ordem $p > 1$ diz-se que tem **convergência supralinear**. Um método que converge com ordem $p = 2$ ou $p = 3$ diz-se que tem **convergência linear, quadrática** ou **cúbica**, respetivamente.

Nota 6.24 (Aproximação numérica da ordem de convergência). Nem sempre é possível encontrar o valor analítico da ordem de convergência. Nesse caso, pode-se tentar estimar numericamente o seu valor. Note-se que para erros pequenos temos

$$|e_{n+1}| \approx C|e_n|^p$$

logo aplicando o logaritmo temos

$$\log|e_{n+1}| \approx p\log|e_n| + \log C.$$

Assim, dividindo ambos os membros por $\log|e_n|$ temos a aproximação numérica para a ordem de convergência

$$p \approx \frac{\log|e_{n+1}|}{\log|e_n|} \tag{6.7}$$

válida para erros pequenos, uma vez que se o método convergir, temos

$$\lim_{n\to\infty} \frac{\log C}{\log|e_n|} = 0.$$

No entanto, para fazermos o calculo numérico precisamos de conseguir calcular o erro e logo de conhecer a solução exata x, o que não é geralmente o caso.

Como ilustrámos no final da secção 6.1, o método da bisseção não garante que o erro em iteradas sucessivas diminua. No entanto, em média a ordem de convergência do método da bisseção é linear, como comprova a fórmula de erro (6.2), que nos indica que o erro passa sensivelmente para metade com o decorrer das iterações, isto é, $p = 1$ e $C = 1/2$.

A partir da definição anterior é também fácil mostrar que o método do ponto fixo tem pelo menos convergência linear.

Teorema 6.25 (Ordem de convergência do método do ponto fixo).

Em geral, nas condições de convergência do teorema 6.16, o método do ponto fixo tem ordem de convergência linear.

Demonstração. Por definição de ponto fixo, temos

$$|x_{n+1} - x| = |g(x_n) - g(x)| \leq L|x_n - x|, \tag{6.8}$$

o que termina a demonstração. □

Convém também notar que em circunstâncias especiais, o método do ponto fixo pode ter convergência superior. Temos o seguinte resultado.

Teorema 6.26 (Ordem de convergência superior do método do ponto fixo).
Seja $g \in C^p(I)$ tal que o método do ponto fixo $x_{n+1} = g(x_n)$ é convergente para o ponto fixo x_g de g em I. Suponha-se ainda que

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

Então o método do ponto fixo tem ordem de convergência p para o ponto fixo x_g.

Demonstração. Pelo desenvolvimento em Fórmula de Taylor (5.2) de g em torno de x_g temos para qualquer $x \in I$, que existe um $\xi \in I$ tal que

$$g(x) = g(x_g) + g'(x_g)(x - x_g) + \cdots + \frac{g^{(p-1)}(x_g)}{(p-1)!}(x - x_g)^{p-1} + \frac{g^{(p)}(\xi)}{p!}(x - x_g)^p,$$

de onde se obtem pela hipótese que

$$g(x) - g(x_g) = \frac{g^{(p)}(\xi)}{p!}(x - x_g)^p.$$

Assim temos para as iteradas x_n do método do ponto fixo que

$$|x_{n+1} - x_g| = |g(x_n) - g(x_g)| = \frac{|g^{(p)}(\xi)|}{p!}|x - x_g|^p \leq \frac{\max_{x \in I} |g^{(p)}(x)|}{p!}|x - x_g|^p$$

o que conclui a demonstração. □

Deixamos para o final do capítulo um exemplo de ordem de convergência superior, nomeadamente no exercício 6.48.

6.2.1 Método do ponto fixo generalizado

Ao contrário do método da bisseção, o método do ponto fixo é facilmente generalizável para contextos de equações não-lineares gerais. Assim sendo, vamos formular os resultados para a resolução de equações

$$Ax = x$$

em que o operador $A : X \to X$ que opera no espaço normado X pode não ser necessariamente linear. A necessidade de X ser um espaço normado nesta formulação, tem apenas a ver com o facto de ser necessário ter uma norma para o operador A. De forma semelhante temos a definição de operador contração.

Definição 6.27 (Contração).

Seja X um espaço normado e U um seu subconjunto.

Diz-se que o operador $A : U \to X$ é uma **contração** se existir uma constante $0 \leq L < 1$ tal que

$$\|Ax - Ay\| \leq L\|x - y\| \tag{6.9}$$

para qualquer $x, y \in U$.

Da mesma forma que para funções em $\mathbb{R}$, temos que qualquer contração é um operador contínuo, pois se $x_n \to x$, $n \to \infty$, temos

$$\|Ax - Ax_n\| \leq L\|x - x_n\| \to 0, \quad n \to \infty.$$

Da mesma forma, um operador que seja uma contração pode ser visto como um operador de Lipschitz que verifica

$$\|Ax - Ay\| \leq L\|x - y\|$$

com constante de Lipschitz $L < 1$. Podemos também definir ponto fixo de forma semelhante.

Definição 6.28 (Ponto Fixo).

Seja X um espaço normado. Diz-se que $x \in U \subset X$ é **ponto fixo** de $A : U \to X$ se $Ax = x$.

Temos então o seguinte resultado em tudo semelhante ao obtido no caso de funções reais.

Teorema 6.29 (Ponto Fixo).

Seja X um espaço normado e U um seu subconjunto completo. Seja ainda $A : U \to U$ uma contração com constante de contração $L < 1$. Então A tem um único ponto fixo.

Demonstração. A demonstração é semelhante à do teorema 6.15, substituindo $g(x)$ por Ax, I por U e as distâncias $|.|$ pela norma considerada $\|.\|$. □

Também no caso generalizado, podemos aplicar o método do ponto fixo, definido de seguida neste contexto generalizado.

Teorema 6.30 (Método do Ponto Fixo).

*Nas condições do teorema 6.29, o **método do ponto fixo** definido por uma iterada inicial $x_0 \in U$ e pela iteração*

$$x_{n+1} = Ax_n$$

para $n \in \mathbb{N}_0$ converge para o único ponto fixo de A em U. Mais ainda, temos as estimativas de erro a priori

$$\|x - x_n\| \leq \frac{L^n}{1-L}\|x_1 - x_0\| \tag{6.10}$$

e a posteriori

$$\|x - x_n\| \leq \frac{L}{1-L}\|x_n - x_{n-1}\|. \tag{6.11}$$

.

Demonstração. Semelhante à demonstração do teorema 6.16. □

A estimativa *a posteriori* (6.11) é tão boa ou melhor que a estimativa *a priori* (6.10), uma vez que L é um majorante para o ganho em termos de erro de uma iteração.

Exercício 6.31. [1] Considere o espaço completo de funções de quadrado integrável

$$L^2([0,1]) = \{f : [0,1] \to \mathbb{R} : \|f\|_2 < \infty\}$$

com $\|f\|_2 = \sqrt{\int_0^1 [f(x)]^2 dx}$. Considere ainda o operador $A : L^2([0,1]) \to L^2([0,1])$ definido por

$$(Af)(x) = \left(x - \frac{1}{2}\right) f(x).$$

(a) Mostre que de facto $A(L^2([0,1])) \subset L^2([0,1])$.

(b) Mostre que A é uma contração.

(c) Calcule a segunda iteração pelo método do ponto fixo, começando com a iterada inicial $f_0(x) = x^2$.

(d) Calcule as estimativas *a priori* e *a posteriori* para a aproximação da alínea.

(e) Calcule uma iteração pelo método do ponto fixo, começando com a iterada inicial $f_0(x) = 0$.

(f) Que conclui?

Resposta.

[1] Este exercício requer alguma familiaridade com Análise Funcional, em particular, operadores integrais entre espaços de funções. Se o leitor não estiver nesse caso, poderá ignorar este exercício.

(a) Temos

$$\|Af\|_2^2 = \int_0^1 \left[\left(x - \frac{1}{2}\right) f(x)\right]^2 dx \leq \max_{x\in[0,1]} \left(x - \frac{1}{2}\right)^2 \|f\|_2^2 \leq \frac{1}{4}\|f\|_2^2$$

logo se $\|f\|_2 \leq \infty$ temos $\|Af\|_2 \leq \infty$ como queríamos demonstrar.

(b) Temos

$$\|Af - Ag\|_2^2 = \int_0^1 \left[\left(x - \frac{1}{2}\right) (f(x) - g(x))\right]^2 dx \leq \max_{x\in[0,1]} \left(x - \frac{1}{2}\right)^2 \|f - g\|_2^2$$

logo A é uma contração com $L = \sqrt{\max_{x\in[0,1]} \left(x - \frac{1}{2}\right)^2} = 1/2$.

(c) Temos

$$f_1(x) = (Af_0)(x) = x^3 - \frac{x^2}{2};$$
$$f_2(x) = (Af_1)(x) = x^4 - x^3 + \frac{x^2}{4};$$

Os gráficos da iterações podem ser vistos na figura 6.1.

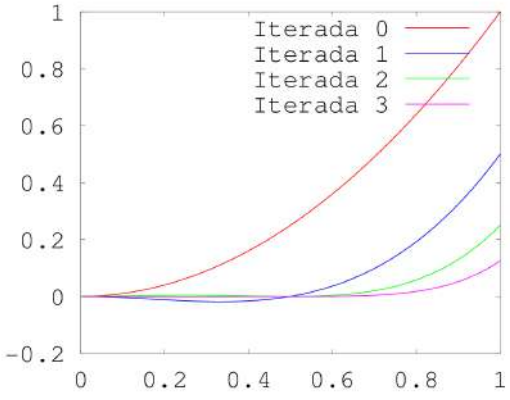


Figura 6.1: *Três primeiras iterações pelo método do ponto fixo referentes ao exercício 6.31(c).*

(d) De (6.10) temos a estimativa para o erro em relação ao ponto fixo f dada por

$$\|f - f_2\|_2 \leq \frac{(1/2)^2}{1-1/2}\|f_1 - f_0\|_2 = \frac{0.30472}{2} \approx 0.15236,$$

uma vez que $\|f_1 - f_0\|_2 = \sqrt{1/7 - 1/2 + 9/20} \approx 0.30472$. Por (6.11) temos a estimativa *a posteriori*

$$\|f - f_2\|_2 \leq \frac{1/2}{1-1/2}\|f_2 - f_1\|_2 \leq 0.096568$$

uma vez que $\|f_2 - f_1\|_2 = \sqrt{1/9 - 1/2 + 11/14 - 1/2 + 9/80} \approx 0.096568$.

(e) Temos

$$f_1(x) = (Af_0)(x) = 0.$$

(f) Pelas alíneas (a) e (b) e do facto de $L^2([0,1])$ ser um espaço completo concluímos que A tem um único ponto fixo e que a iteração do método do ponto fixo $f_{n+1} = Af_n$ converge para o ponto fixo de A para qualquer iterada inicial $f_0 \in L^2([0,1])$. Da alínea anterior, concluímos que $Af = f$ para $f(x) = 0$ logo a função identicamente nula é o único ponto fixo de A. As iteradas da alínea (c) parecem de facto tender para a função identicamente nula.

6.3 Método de Newton

Como vimos na secção anterior, o método do ponto fixo tem em geral convergência linear. O método de Newton que apresentamos nesta secção é um caso particular de um método do ponto fixo com convergência local quadrática, isto é, com $p = 2$ na condição (6.8). O objetivo do método de Newton é encontrar raízes de dada função f reescrevendo a equação $f(x) = 0$ como

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

ou seja, o problema de obter a raiz de f é equivalente ao problema de encontrar o ponto fixo de g com expressão

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{6.12}$$

A aplicação do método do ponto fixo à equação anterior é chamado o método de Newton.

Definição 6.32 (Método de Newton).

Seja $f \in C^1(I)$. Então chama-se **método de Newton** ao método iterativo

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{6.13}$$

a partir de uma iterada inicial $x_0 \in I$.

Exercício Octave 6.33. Escreva uma função em Octave

`metNewton(expf,expdf, x0, Niter)`

que dadas as expressões `expf` e `expdf` da função f e da sua primeira derivada f', a iterada inicial `x0` determine $Niter$ iterações pelo método de Newton.

Resposta.

Temos a função em Octave:

```
function lx = metNewton(expf,expdf, x0, Niter)
% Calcula niter iteracoes do ponto fixo.
%
% INPUT: expf - expressão da função f
% expdf - expressão da derivada função f
% x0 - iterada inicial
% niter - numero de iteracoes
% L - constante de contracao
%
% OUTPUT: lx - lista da iteradas
f = inline(expf);
df = inline(expdf);
lx = zeros(Niter+1,1);
lx(1) = x0;
for iter = 1:  Niter
  lx(iter+1) = lx(iter)-f(lx(iter))/df(lx(iter));
end
disp(['Iterada ',num2str(Niter),' = ', num2str(lx(end))]);
return
```

A escolha da função (6.12) para a iteração do método não é inocente. De facto consegue-se mostrar com algumas condições adicionais que se o método de Newton é convergente, então converge com ordem quadrática, melhorando portanto a performance em relação ao método do ponto fixo usual.

Teorema 6.34 (Convergência local do método de Newton).

Seja $f \in C^2([a,b])$ tal que $f'(x) \neq 0$ para qualquer $x \in [a,b]$.

Então o método de Newton (6.13) *é localmente convergente para a raiz z de f em I, isto é, se a iterada inicial x_0 for suficientemente próxima da raiz z, o método é convergente. Além disso, o método tem ordem de convergência quadrática , isto é,*

$$|e_{n+1}| \leq C|e_n|^2, \tag{6.14}$$

em que $e_n := z - x_n$ é o erro da iterada n e $C := \dfrac{\max\limits_{x\in[a,b]} |f''(x)|}{2 \min\limits_{x\in[a,b]} |f'(x)|}$, se se verificar a condição local de convergência

$$C|z - x_0| = L_0 < 1. \tag{6.15}$$

Nessas condições, definindo

$$L_n := C|e_n| < 1$$

temos a estimativa de erro a posteriori

$$|e_{n+1}| \leq \frac{C}{(1-L_n)^2}|x_{n+1} - x_n|^2 \tag{6.16}$$

Demonstração. Começamos por notar que pela fórmula de Taylor (5.2) de f em torno de $x_n \in [a,b]$, existe $\xi \in [a,b]$ tal que

$$\underbrace{f(z)}_{=0} = f(x_n) + f'(x_n)e_n + \frac{f''(\xi)}{2}e_n^2$$

e logo dividindo por $f'(x_n) \neq 0$ obtemos

$$z - \left(x_n - \frac{f(x_n)}{f'(x_n)}\right) = -\frac{f''(\xi)}{2f'(x_n)}e_n^2. \tag{6.17}$$

Assim é claro que

$$|e_{n+1}| = |z - x_{n+1}| = \left|z - \left(x_n - \frac{f(x_n)}{f'(x_n)}\right)\right| = \left|\frac{f''(\xi)}{2f'(x_n)}\right| e_n^2.$$

Como $f \in C^2([a,b])$ as suas primeira e segunda derivadas são contínuas em [a,b], logo pelo teorema de Weierstrass 6.9 estas têm máximo e mínimo em $[a,b]$. Assim, como $f'(x) \neq 0$, existe $C > 0$ tal que

$$\left|\frac{f''(\xi)}{2f'(x_n)}\right| \leq \frac{\max\limits_{x\in[a,b]} |f''(x)|}{2 \min\limits_{x\in[a,b]} |f'(x)|} = C, \quad \forall x_n \in [a,b].$$

Assim obtemos

$$|e_{n+1}| \leq C|e_n|^2,$$

logo se o método convergir, converge com convergência quadrática. Para garantir convergência local, basta assegurar que x_0 satisfaz

$$C|z - x_0| = L_0 < 1$$

que equivale a (6.15), de forma a garantir que

$$|e_{n+1}| \leq L_n|e_n| < |e_n|,$$

e logo convergência. Finalmente, para demonstrar a estimativa (6.16), temos

$$e_n = z - x_n = z - x_{n+1} + (x_{n+1} - x_n) = e_{n+1} + (x_{n+1} - x_n)$$

e logo por Cauchy-Schwartz temos

$$|e_n| \leq |e_{n+1}| + |x_{n+1} - x_n|.$$

Tomando agora pela estimativa (6.14) que

$$|e_n| \leq C|e_n|^2 + |x_{n+1} - x_n| \leq L_n|e_n| + |x_{n+1} - x_n|$$

logo

$$(1 - L_n)|e_n| \leq |x_{n+1} - x_n|$$

ou seja

$$|e_n| \leq \frac{1}{1 - L_n}|x_{n+1} - x_n|.$$

A estimativa (6.16) sai agora diretamente da estimativa (6.14), majorando $|e_n|^2$ com o majorante anterior. □

Exercício 6.35. Determine uma aproximação da solução de $x = \sin(x+1)$ no intervalo $[0, 1]$, tomando 16 iterações do método de Newton com iterada inicial $x_0 = 0$ de forma a comparar a sua performance com a do método do ponto fixo e da bisseção na tabela 6.1. Mostre que o método é convergente.

Resposta.

Consideramos

$$f(x) = \sin(x+1) - x$$

uma vez que o método de Newton determina raizes de f. Começamos por mostrar a convergência do método. Temos

$$f'(x) = \cos(x+1) - 1, \quad f''(x) = -\sin(x+1)$$

Como no intervalo $[0,1]$ temos $f'' < 0$, concluímos que f' é decrescente. Assim,

$$\min_{x\in[0,1]} |f'(x)| = \min\{|f'(0)|, |f'(1)|\} \geq 0.45970.$$

Por outro lado, temos

$$\max_{x\in[0,1]} |f''(x)| = \sin(\pi/2) = 1.$$

Assim obtemos que

$$C = \frac{\max\limits_{x\in[a,b]} |f''(x)|}{2 \min\limits_{x\in[a,b]} |f'(x)|} = 1.0877$$

e logo obtemos a condição de convergência

$$|z - x_0| < \frac{1}{C} = 0.9194.$$

Como $x_0 = 0.5$ e $z \in [0,1]$ a condição está automaticamente satisfeita, logo o método de Newton é convergente.

Assim, temos as iteradas da tabela 6.2, obtidas com a função em Octave do exercício 6.33.

É fácil verificar que o método de Newton converge muito mais rapidamente que o do ponto fixo ou da bisseção, aliás, como é esperado pela superior ordem de convergência. Note-se que com 5 iterações do método de Newton o erro já é inferior ao obtido com 32 iterações do método do ponto fixo ou da bisseção (ver tabela 6.1 e 6.2).

O teorema 6.34 apenas mostra convergência local do método de Newton, isto é, garantem convergência caso a iterada inicial x_0 estiver suficientemente próxima da solução z. Mais ainda, a demonstração do teorema 6.34 e em particular a equação (6.15) mostra que se

$$|z - x_0| < \frac{1}{C}, \quad \text{com } C = \frac{\max\limits_{x\in[a,b]} |f''(x)|}{2 \min\limits_{x\in[a,b]} |f'(x)|}$$

n	Met. Newton	
	x_n	Erro
0	0.500000000000000	0.434563
1	1.035365224451990	0.100802
2	0.937750050464278	0.00318684
3	0.934566700543645	3.48979×10^{-6}
4	0.934563210756222	4.19731×10^{-12}
5	0.934563210752024	1.11022×10^{-16}
6	0.934563210752024	$< 10^{-16}$

Tabela 6.2: *Tabela de iterações do método de Newton para determinar a solução de* $x = \sin(x+1)$, *para comparar com a performance do método do ponto fixo e da bisseção na tabela 6.1.*

o método converge. A condição anterior é de difícil verificação, uma vez que implica determinar majorantes e minorantes das derivadas de f no intervalo. Nesta perspetiva, vamos de seguida apresentar outros resultados de convergência de mais fácil verificação que nos indiquem condições para a iterada inicial de forma a que o método de Newton convirja para a solução z.

Teorema 6.36 (Convergência do método de Newton).

Seja $f : [a,b] \to \mathbb{R}$ *uma função duas vezes diferenciável (i.e.,*$f \in C^2([a,b])$*) e seja a iterada inicial* $x_0 \in [a,b]$ *tais que*

(a) $f(a)f(b) < 0$;

(b) $f'(x) \neq 0$ *em* $[a,b]$;

(c) $f''(x) \geq 0$ *em* $[a,b]$ *ou* $f''(x) \leq 0$ *em* $[a,b]$;

(d) $f(x_0).f''(x) \geq 0$;

Então o método de Newton (6.13) *dado por* $x_0 \in [a,b]$ *e*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{6.18}$$

é convergente (monotonamente) com ordem quadrática para a única raiz de f *em* $[a,b]$. *Tiram-se as mesmas conclusões a partir da primeira iterada* x_1 *se se substituir a condição* (d) *por*

(d2) $\left|\frac{f(a)}{f'(a)}\right| < b-a$ *e* $\left|\frac{f(b)}{f'(b)}\right| < b-a$;

Demonstração. Pelo corolário 6.7, as condições (a) e (b) garantem a existência de uma única raiz z de f em $[a, b]$. Vamos assumir que $f' > 0$ e $f'' \geq 0$, sendo que nos restantes casos o procedimento de prova é semelhante. Assim f é crescente, pelo que $f(a) < 0$ e $f(b) > 0$. Para que a condição (d) seja satisfeita, temos que $f(x_0) \geq 0$ e logo $x_0 \in [z, b]$. Temos então que

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

logo concluímos que $x_1 \leq x_0$. Por outro lado, de (6.17) temos que existe $\xi \in [z; x_0]$ tal que

$$z - x_1 = -\frac{f''(\xi)}{2f'(x_n)}(z - x_n)^2 \leq 0$$

logo concluímos que $z \leq x_1 \leq x_0$. Assim, repetindo o processo para x_n provamos por indução que

$$x_0 \geq x_1 \geq \cdots \geq x_n \geq \cdots \geq z.$$

Assim a sucessão $(x_n)_n \in \mathbb{N}$ é limitada e monótona, logo é convergente [7]. Sendo x o limite dessa sucessão, temos por continuidade de g dada por (6.12) que

$$x = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g\left(\lim_{n\to\infty} x_n\right) = g(x)$$

logo como g tem como único ponto fixo em $[a, b]$ a raiz z de f, temos que $x_n \to z$, logo convergência.

Falta apenas mostrar que se $(d2)$ for satisfeita em vez de (d) então x_1 satisfaz (d). Supomos então $f(x_0) < 0$, ou seja, que $x_0 \in [a, z]$, pois no outro caso já está provado. Temos também que

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2} \leq 0, \quad x \in [a, b]$$

logo $g'(x)$ é decrescente nos intervalo $[a, z]$. Assim

$$x_0 < z \quad \Rightarrow \quad g(x_0) > g(z)$$

logo concluímos que $x_1 = g(x_0) > g(z) = z$, ou seja $x_1 \in [z, b]$ e logo $f(x_1) > 0$, como queríamos demonstrar.

□

Exercício 6.37. Considere a equação

$$x^2 = \cos(x).$$

(a) Mostre que a equação tem uma única solução em $I = [1/2, 1]$.

(b) Defina uma função f e uma iterada inicial $x_0 \in I$ tal que o método de Newton convirja para a solução da equação.

(c) Calcule uma aproximação para a solução da equação com erro menor que 10^{-8}.

Resolução.

(a) Seja $f(x) = x^2 - \cos(x)$. Pretendemos mostrar que f tem apenas uma raiz no intervalo I. Assim, como $f'(x) = 2x + \sin(x) > 0$ (uma vez que $\sin(x) > 0$ para $x \in I$) e $f(1/2).f(1) = -0.28850 < 0$, temos pelo corolário 6.7 que f tem uma única raiz em I e logo que a equação tem uma única solução neste intervalo.

(b) Definimos f como na alínea anterior. Tendo em atenção o teorema 6.36, as condições (a) e (b) já foram verificadas na alínea anterior. Temos também que $f''(x) = 2 + \cos(x) > 0$ (pois $-1 \leq \cos(x) \leq 1$) logo a condição (c) também está verificada. Pela condição (d) do teorema 6.36, basta encontrar uma iterada inicial x_0 tal que $f(x_0) > 0$, ou seja, por exemplo $x_0 = 1$. Por outro lado, como

$$\left|\frac{f(1/2)}{f'(1/2)}\right| = 0.424 < 0.5, \quad \left|\frac{f(1)}{f'(1)}\right| = 0.162 < 0.5,$$

a condição $(d2)$ é verificada, logo temos pelo teorema 6.36 que o método de Newton é convergente para qualquer iterada inicial $x_0 \in I$.

(c) Como $f''(x)$ é decrescente e positiva em I e $f'(x)$ é crescente e positiva em I, temos

$$C = \frac{\max\limits_{x\in[a,b]} |f''(x)|}{2 \min\limits_{x\in[a,b]} |f'(x)|} = \frac{f''(1/2)}{2f'(1/2)} = \frac{2.878}{2 \times 1.479} = 0.973$$

Assim, para $x_0 = 1$, temos $|z - x_0| < 1/2$ e logo $C|z - x_0| \leq C/2 =: L_0 < 1$, logo é válida a estimativa de erro (6.16). Assim, com os comandos Octave:

```
Niter= 4;
expf = 'x^2-cos(x)';
expdf = '2*x+sin(x)';
x0=1;
C= (2+cos(1/2))/2/(1+sin(1/2));
lx= metNewton(expf,expdf, x0, Niter);
A=zeros(size(lx,1)-1,2);
```

```
A(:,1)=lx(2:end);
erroant =0.5;
for ii =1:Niter
  Ln=C*erroant;
  A(ii,2) = C*abs(lx(ii+1)-lx(ii))^2/(1-Ln)^2;
  erroant =A(ii,2);
end
```

obtemos a seguinte tabela:

n	x_n	$\|z - x_n\| \leq C\|x_n - x_{n-1}\|^2/(1 - L_n)^2 = \ldots$
1	0.838218409904707	0.0964468
2	0.824241868225874	0.000231342
3	0.824132319050929	1.16767×10^{-8}
4	0.824132312302522	4.42902×10^{-17}

logo são necessárias $n = 4$ para garantir um erro inferior a 10^{-8}.

6.3.1 Método de Newton Generalizado

O método de Newton pode também ser aplicado num contexto mais geral, nomeadamente no caso em que $F : X \to Y$ é um operador entre espaços normados. No entanto, nesse caso é preciso definir a derivada do operador F. Temos então a seguinte definição.

Definição 6.38 (Derivada de Frechet).

Seja $F : X \to Y$ um operador limitado entre espaços normados. Diz-se que F é um operador **Frechet-diferenciável** no ponto $x_0 \in X$ se existir um operador limitado $F'(x_0) : X \to Y$ tal que

$$\|F(x_0 + h) - F(x_0) - [F'(x_0)](h)\| = o(\|h\|), \quad \|h\| \to 0, \tag{6.19}$$

ou seja, o limite seguinte existe e

$$\lim_{\|h\|\to 0} \frac{\|F(x_0 + h) - F(x_0) - [F'(x_0)](h)\|}{\|h\|} = 0 \tag{6.20}$$

Se F é Frechet-diferenciável em todos os pontos de X então F diz-se **Frechet-diferenciável em X.**

Convém fazer algumas notas em relação à derivação de Frechet.

Nota 6.39. A primeira é que se F é Frechet-diferenciável em x, então F é contínuo em x. Basta verificar que se $x_n \to x$ então

$$\|F(x_n) - F(x)\| - \|[F'(x)](x_n - x)\| \leq o(\|x_n - x\|)$$

pelo facto que $\|a - b\| \geq \|a\| - \|b\|$, logo

$$\|F(x_n) - F(x)\| \leq \|[F'(x)]\|\|(x_n - x)\| + o(\|x_n - x\|) \to 0,$$

pela desigualdade de Schwartz.

Nota 6.40. Note-se também que a derivada de Frechet é única. De facto, se existissem duas derivadas de Frechet $F_1'(x_0)$ e $F_2'(x_0)$ do operador F em x_0, teríamos de forma semelhante que

$$\begin{aligned}\|[F_1'(x_0)]h - [F_2'(x_0)]h\| &\leq \Big\|F(x_0 + h) - F(x_0) - \Big(F(x_0 + h) - F(x_0)\Big)\Big\| + o(\|h\|) \\ &= o(\|h\|), \quad \|h\| \to 0,\end{aligned}$$

logo

$$\begin{aligned}\|[F_1'(x_0)] - [F_2'(x_0)]\| &= \max_{x \in X} \frac{\|[F_1'(x_0)]x - [F_2'(x_0)]x\|}{\|x\|} \\ &\leq \lim_{\|h\| \to 0} \frac{\|[F_1'(x_0)]h - [F_2'(x_0)]h\|}{\|h\|} \\ &= 0,\end{aligned}$$

ou seja, as duas derivadas coincidem.

Nota 6.41. Se F é uma função de $\mathbb{R}^n$ em $\mathbb{R}^n$ então a derivada de Frechet coincide com o Jacobiano de F, saindo a prova da fórmula de Taylor. Assim, $F'(x)$ é uma matriz $n \times n$ cuja entrada na posição (i, j) é dada por

$$[F'(x)]_{i,j} = \frac{\partial F_i}{\partial x_j}(x).$$

Temos então a formulação do método de Newton Generalizado.

Teorema 6.42 (Método de Newton generalizado).

Seja $F : X \to Y$ um operador limitado e Frechet-diferenciável entre espaços normados. Então o método de Newton Generalizado é definido por uma iterada inicial x_0 e as iteradas seguintes

$$x_{n+1} = x_n - [F'(x_n)]^{-1} F(x_n). \tag{6.21}$$

Se a derivada de Fréchet F' for bem condicionada em torno da raiz z, então o método converge localmente (isto é, dada uma iterada inicial x_0 suficientemente próxima de z) com ordem quadrática para solução z de

$$F(x) = 0.$$

Demonstração. Não faremos a prova, ainda que a ideia seja semelhante ao caso do método de Newton para funções reais. Ver [5]. □

Nota 6.43. Normalmente, cada iterada do método de Newton passa por resolver um sistema linear. Na realidade, a iteração (6.21) pode ser reescrita como

$$\begin{cases} [F'(x_n)]h = -F(x_n), \\ x_{n+1} = x_n + h, \end{cases}$$

em que o primeiro passo envolve resolver o sistema linear para obter a atualização h e o segundo passo é apenas atualizar para a nova iterada.

Exercício Octave 6.44.

Pretende-se determinar a interseção para $x > 0$ entre a circunferência de raio unitário e o gráfico da função $\sin(x)$. Determine três iterações do método de Newton para o efeito.

Resolução.

Pretendemos obter o ponto (x, y) com $x > 0$ que satisfaz simultaneamente

$$\begin{cases} x^2 + y^2 = 1, \\ y = \sin(x). \end{cases}$$

Assim, pretendemos encontrar o zero da função $F : \mathbb{R}^2 \to \mathbb{R}^2$ com expressão

$$F(x, y) = \begin{bmatrix} x^2 + y^2 - 1 \\ y - \sin(x) \end{bmatrix}.$$

Tomando em atenção a nota 6.41, temos

$$F'(x, y) = \begin{bmatrix} 2x & 2y \\ -\cos(x) & 1 \end{bmatrix}.$$

Assim, partindo da iterada inicial $x_0 = [1, 0]$ (escolhida pelo facto de pretendermos $x > 0$), na primeira iterada que

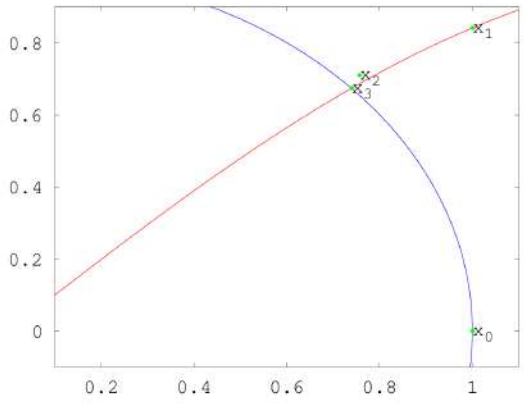
$$\begin{aligned} x_1 &= x_0 + [F'(x_0)]^{-1} F(x_0) \\ &= \begin{bmatrix} 1^2 + 0^2 - 1 \\ 0 - \sin(1) \end{bmatrix} - \begin{bmatrix} 2 \times 1 & 2 \times 1 \\ -\cos(1) & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1^2 + 0^2 - 1 \\ 0 - \sin(1) \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0.84147 \end{bmatrix} \end{aligned}$$

Assim, com a rotina em Octave dada por

```
Niter=4;
F=inline('[x^2+y^2-1; y - sin(x)]','x','y');
DF=inline('[2*x,2*y; -cos(x),1]','x','y');
lx =zeros(2,Niter+1);
lx(:,1)=[1,0].';
for ii=1:Niter
  lx(:,ii+1)= lx(:,ii) - ...
    inv(DF(lx(1,ii),lx(2,ii)))*F(lx(1,ii),lx(2,ii));
end
```

obtemos as três primeiras iterações:

	x	y
x_0	1.0000000000	0.0000000000
x_1	1.0000000000	0.8414709848
x_2	0.7566170403	0.7099706105
x_3	0.7396666658	0.6741397407



6.4 Método da Secante

O método da secante é um outro método para determinar raízes da função f, que surge como uma necessidade de adaptar o método de Newton no caso em que a função f não é diferenciável ou não se conhece, é muito complicado ou computacionalmente custoso o cálculo da sua derivada. Pela Fórmula de Taylor (5.2) temos que

$$f(x_{n-1}) = f(x_n) + f'(x_n)(x_{n-1} - x_n) + O(|x_n - x_{n-1}|^2)$$

pelo que temos a aproximação para a derivada

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

com um erro da ordem de $O(|x_n - x_{n-1}|)$. Desta forma, na equação (6.13) substituise o valor de $f'(x_n)$ pela aproximação anterior, originado assim o método da secante com expressão iterativa

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

De notar portanto que para determinar o valor de uma nova iterada pelo método da secante é necessário o valor das duas iteradas anteriores. Em particular, para iniciar o método da secante são necessárias duas iteradas iniciais x_{-1} e x_0.

Teorema 6.45 (Método da Secante).
Seja $f : [a, b] \to \mathbb{R}$ uma função em $C^2([a, b])$.
Então o método da secante com iteradas iniciais $x_{-1}, x_0 \in [a, b]$ e iteradas seguintes

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \tag{6.22}$$

é localmente convergente, isto é, converge se as iteradas inciais estiverem suficientemente próximas da raiz de z de f. Nessas condições, temos que existe um $C > 0$ tal que

$$|e_{n+1}| \leq C|e_n||e_{n-1}|, \tag{6.23}$$

em que $e_n = z - x_n$ é o erro da iterada n. Mais ainda, temos o limite

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n e_{n-1}} = -\frac{f''(z)}{2f'(z)}. \tag{6.24}$$

Demonstração. Vamos mostrar apenas o limite (6.24), uma vez que este implica que localmente a estimativa (6.23) é válida, e esta implica que para iteradas inciais x_{-1} e x_0 suficientemente próximas da raiz z, o erro diminui e logo o método é localmente convergente.

Começamos por considerar que pela expressão (6.22), obtemos

$$\begin{aligned} e_{n+1} &= x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} - z \\ &= \frac{-f(x_n)(x_n - x_{n-1}) + [f(x_n) - f(x_{n-1})](x_n - z)}{f(x_n) - f(x_{n-1})} \\ &= \frac{f(x_n)(x_{n-1} - z) - f(x_{n-1})(x_n - z)}{f(x_n) - f(x_{n-1})} \\ &= \left[\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\right]\left[\frac{f(x_n)e_{n-1} - f(x_{n-1})e_n}{x_n - x_{n-1}}\right]. \end{aligned} \tag{6.25}$$

Pela fórmula de Taylor de primeira ordem, temos que existe $\xi_1 \in [x_n; x_{n-1}]$ tal que

$$f(x_n) = f(x_{n-1}) + f'(\xi_1)(x_n - x_{n-1})$$

logo

$$f(\xi_1) = \frac{f(x_n) - f(x_n)}{x_n - x_{n-1}}. \tag{6.26}$$

Por outro lado, temos que pela fórmula de Taylor de quarta ordem que existe $\xi_2 \in [x_n; z]$ tal que

$$f(x_n) = f(z) + f'(z)(x_n - z) + \frac{f''(z)}{2}(x_n - z)^2 - \frac{f'''(z)}{6}(x_n - z)^3 + \frac{f^{(4)}(\xi_2)}{24}(x_n - z)^4,$$

logo

$$f(x_n) = -f'(z)e_n + \frac{f''(z)}{2}e_n^2 - \frac{f'''(\xi_3)}{6}e_n^3 + \frac{f^{(4)}(\xi_2)}{24}e_n^4$$

e similarmente que existe $\xi_3 \in [x_{n-1}; z]$ tal que

$$f(x_{n-1}) = -f'(z)e_{n-1} + \frac{f''(z)}{2}e_{n-1}^2 - \frac{f'''(z)}{6}e_{n-1}^3 + \frac{f^{(4)}(\xi_3)}{24}e_{n-1}^4.$$

Assim, como $x_n - x_{n-1} = -(e_n - e_{n+1})$ concluímos que

$$\begin{aligned}
&\frac{f(x_n)e_{n-1} - f(x_{n-1})e_n}{x_n - x_{n-1}} = \\
&\frac{-\frac{f''(z)}{2}(e_n - e_{n+1}) + \frac{f'''(z)}{6}(e_n^2 - e_{n-1}^2) + \frac{f^{(4)}(\xi_2)}{24}e_n^4 - \frac{f^{(4)}(\xi_3)}{24}e_{n-1}^4}{e_n - e_{n+1}}e_n e_{n-1} \\
&= -\frac{f''(z)}{2}e_n e_{n-1} + \frac{f'''(z)}{6}(e_n + e_{n-1})e_n e_{n-1} + \frac{\frac{f^{(4)}(\xi_2)}{24}e_n^4 - \frac{f^{(4)}(\xi_3)}{24}e_{n-1}^4}{e_n - e_{n+1}}e_n e_{n-1} \\
&= -\frac{f''(z)}{2}e_n e_{n-1} + O\Big((e_n + e_{n-1})e_n e_{n-1}\Big).
\end{aligned} \tag{6.27}$$

para e_n e e_{n-1} pequenos, ou seja, para iteradas próximas da raiz z. Assim, se e_n e e_{n-1} forem suficientemente pequenos, temos de (6.25)-(6.27), que

$$e_{n+1} = -\frac{f''(z)}{2f'(\xi_1)}e_n e_{n-1} + O\Big((e_n + e_{n-1})e_n e_{n-1}\Big)$$

o que demostra o resultado uma vez que

$$\frac{e_{n+1}}{e_n e_{n-1}} = -\frac{f''(z)}{2f'(\xi_1)} + O\Big(e_n + e_{n-1}\Big).$$

□

Uma vez que o método da secante é uma aproximação do método de Newton, consegue-se mostrar a convergência do método da secante se forem satisfeitas as condições (a) a (c) do teorema 6.36 com a condição (d) substituída por

(d) $f(x_0).f''(x) > 0$ e $f(x_1).f''(x) > 0$

uma vez que o método da secante tem duas iteradas iniciais.

Vamos agora mostrar que o método da secante tem convergência supralinear.

Teorema 6.46 (Ordem de Convergência do Método da Secante).

Nas condições de convergência, o método da secante tem ordem de convergência supralinear, com $p = \frac{1+\sqrt{5}}{2} \approx 1.618$.

Demonstração. Queremos determinar p tal que para n suficientemente grande temos

$$|z - x_{n+1}| \approx A|z - x_n|^p,$$

com $A > 0$ constante. De (6.24), temos que

$$\begin{aligned} C &= \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n||z - x_{n-1}|} \\ &= \lim_{n\to\infty} \frac{A|z - x_n|^p}{|z - x_n|A^{-1/p}|z - x_n|^{1/p}} \\ &= \tilde{C} \lim_{n\to\infty} |z - x_n|^{p-1-1/p}. \end{aligned}$$

para $\tilde{C} = A^{1+1/p}$, logo concluímos que $p - 1 - 1/p = 0$ de forma a que o limite possa ser constante. Pela fórmula resolvente de segundo grau, obtemos que

$$p - 1 - 1/p = 0 \Leftrightarrow p^2 - p - 1 = 0 \Leftrightarrow p = \frac{1 \pm \sqrt{5}}{2},$$

e escolhendo a raiz positiva, temos o resultado. □

O resultado anterior mostra que o método da secante converge mais lentamente que o de Newton, mas mais rapidamente que o do ponto fixo (no caso geral).

Exercício 6.47. Pretende-se determinar uma aproximação de e, determinando para isso uma aproximação da raiz de $f(x) = 1 - \ln(x)$ no intervalo $I = [2, 3]$.

(a) Mostre que f tem uma única raiz $z \in I$.

(b) Determine 5 iterações pelo método da bisseção e os respetivos erros na aproximação.

(c) Mostre que os pontos fixos de $g_1(x) = x + 1 - \ln(x)$ e $g_2(x) = x - 1 + \ln(x)$ no intervalo I coincidem com a raiz de f.

(d) Mostre que a função g_1 satisfaz as condições de convergência método do ponto fixo.

(e) Mostre que a função g_2 não é uma contração.

(f) Determine 5 iterações pelo do ponto fixo aplicado a g_1 com $x_0 = 2$ e os respetivos erros na aproximação.

(g) Mostre que o método de Newton aplicado a f converge com $x_0 = 2$.

(h) Determine 5 iterações pelo de Newton aplicado a f com $x_0 = 2$ e os respetivos erros na aproximação.

(i) Mostre que o método da Secante aplicado a f converge com $x_{-1} = 2.1$ e $x_0 = 2$.

(j) Determine 5 iterações pelo da secante aplicado a f com $x_{-1} = 2.1$ e $x_0 = 2$ e os respetivos erros nas aproximações.

(k) Determine numericamente a ordem de convergência em cada caso através da expressão (6.7) para as iteradas calculadas.

Resposta.

(a) Uma vez que $f(2).f(3) \approx -0.030 < 0$ e de $f'(x) = -1/x < 0$ para $x \in I$, temos o resultado pelo corolário 6.7.

(b) Ver tabelas 6.3 e 6.4.

(c) Como $g(z) = z \pm f(z) = z$, o resultado está provado.

(d) Temos $g_1'(x) = 1 - 1/x$. Assim concluímos que

$$|g_1'(x)| < g_1'(3) = 2/3 < 1, \forall x \in I.$$

Por outro lado, como $g_1'(x) > 0$, temos que g_1 é crescente e logo

$$g_1(I) \subset [g_1(2), g_1(3)] = [2.3069, 2.9014] \subset I.$$

Assim, pelo corolário 6.17 temos o resultado.

n	Aproximação x_n			
	Bisseção	Ponto Fixo	Newton	Secante
-1	-	-	-	2.1000000000
0	2.5000000000	2.0000000000	2.0000000000	2.0000000000
1	2.7500000000	2.3068528194	2.6137056389	2.6289235231
2	2.6250000000	2.4709686397	2.7162439264	2.7058072617
3	2.6875000000	2.5663584041	2.7182810644	2.7180754903
4	2.7187500000	2.6238704734	2.7182818285	2.7182813546
5	2.7031250000	2.6592199658	2.7182818285	2.7182818284

Tabela 6.3: *Exercício 6.47: Aproximações de $z = e$ pelos 4 métodos.*

n	Erro: $\lvert z - x_n \rvert$			
	Bisseção	Ponto Fixo	Newton	Secante
0	0.218282	0.718282	0.718282	0.718282
1	0.0317182	0.411429	0.104576	0.0893583
2	0.0932818	0.247313	0.0020379	0.0124746
3	0.0307818	0.151923	7.64101×10^{-7}	0.000206338
4	0.000468172	0.0944114	1.07025×10^{-13}	4.73826×10^{-7}
5	0.0151568	0.0590619	4.44089×10^{-16}	1.79843×10^{-11}

Tabela 6.4: *Exercício 6.47: Erros das aproximações de $z = e$ pelos 4 métodos.*

(e) Como $g_2'(x) = 1 + 1/x > 1$, pelo teorema de Lagrange 6.4 temos que g_2 não é uma contração.

(f) Ver tabelas 6.3 e 6.4.

(g) Tendo em atenção o teorema 6.36, as condições (a) e (b) já foram verificadas na alínea (a). Temos também que $f''(x) = 1/x^2 > 0$, logo a condição (c) também é satisfeita. Finalmente, temos que $f(x_0) = f(2) = 0.306 > 0$, logo a condição (d) é satisfeita e o método de Newton é convergente.

(h) Ver tabelas 6.3 e 6.4.

(i) Adicionalmente à alínea (g), falta mostrar que $f(x_{-1})f'' > 0,$ o que se verifica pois $f(2.1) = 0.258 > 0$.

(j) Ver tabelas 6.3 e 6.4.

(k) Temos a tabela 6.5, obtida a partir da expressão (6.7). Verificamos que a ordem de convergência numérica do método da bisseção oscila, como es-

perado, Por outro lado, a do método do ponto fixo parece tender para 1, a do método de Newton para 2 e a da secante para 1.618, os valores teóricos encontrados. O logaritmo do erro na última iteração do método de Newton já entra na precisão da máquina (10^{-16}), pelo que o valor da aproximação numérica da ordem de convergência é dispar dos restantes.

n	Ordem conv. numérica			
	Bisseção	Ponto Fixo	Newton	Secante
1	2.26737	2.684	6.82347	7.29873
2	0.687401	1.5731	2.74414	1.81527
3	1.46739	1.34878	2.27323	1.93565
4	2.20254	1.25245	2.12046	1.71605
5	0.54643	1.19875	1.18365	1.699

Tabela 6.5: *Exercício 6.47: Aproximação numérica* (6.7) *para a ordem de convergência dos 4 métodos.*

Exercício 6.48. Pretende-se determinar uma aproximação de π, determinando para isso uma aproximação da raiz de $f(x) = \sin(x)$ no intervalo $I = [3, 4]$.

(a) Mostre que f tem uma única raiz $z \in I$.

(b) Determine 3 iterações pelo método da bisseção e os respetivos erros na aproximação.

(c) Mostre que os pontos fixos das funções $g_1(x) = x+\sin(x)$ e $g_2(x) = x-\sin(x)$ no intervalo I coincidem com a raiz de f.

(d) Mostre que a função g_1 satisfaz as condições de convergência método do ponto fixo.

(e) Mostre que a função g_2 não é uma contração.

(f) Determine 3 iterações pelo do ponto fixo aplicado a g_1 com $x_0 = 3$ e os respetivos erros na aproximação.

(g) Mostre que o método de Newton aplicado a f converge com $x_0 = 3$.

(h) Determine 3 iterações pelo de Newton aplicado a f com $x_0 = 3$ e os respetivos erros na aproximação.

(i) Determine 3 iterações pelo da secante aplicado a f com $x_{-1} = 3$ e $x_0 = 3.1$ e os respetivos erros nas aproximações.

(j) Como justifica que o método do ponto fixo tenha uma ordem de convergência semelhante ao método de Newton e superior ao da Secante neste caso?

(k) Determine numericamente a ordem de convergência em cada caso através da expressão (6.7) para as iteradas calculadas.

Resposta.

(a) Temos $f(3).f(4) < 0$ e $f'(x) = \cos x < 0$ para $x \in [3, 4] \subset]\pi/2, 3\pi/2[$, logo pelo corolário 6.7 temos o resultado.

(b) Ver tabelas 6.6 e 6.7.

n	Aproximação x_n			
	Bisseção	Ponto Fixo	Newton	Secante
-1	-	-	-	3.1000000000
0	3.5000000000	3.0000000000	3.0000000000	3.0000000000
1	3.2500000000	3.1411200081	3.1425465431	3.1417730920
2	3.1250000000	3.1415926536	3.1415926533	3.1415920500
3	3.1875000000	3.1415926536	3.1415926536	3.1415926536

Tabela 6.6: *Exercício 6.48: Aproximações de $z = \pi$ pelos 4 métodos.*

n	Erro: $\lvert z - x_n \rvert$			
	Bisseção	Ponto Fixo	Newton	Secante
0	0.358407	0.141593	0.141593	0.141593
1	0.108407	0.000472646	0.000953889	0.000180438
2	0.0165927	1.75975×10^{-11}	2.893174×10^{-10}	6.0356×10^{-7}
3	0.0459073	$< 10^{-16}$	$< 10^{-16}$	3.10862×10^{-15}

Tabela 6.7: *Exercício 6.48: Erros das aproximações de $z = \pi$ pelos 4 métodos.*

(c) Para a raiz z de $f(x) = \sin(x)$ temos que $g(z) = z \pm \sin(z) = z$, logo o resultado está provado.

(d) Temos $g_1'(x) = 1 + \cos x$ e $g_1''(x) = -\sin(x)$, logo g_1' é decrescente em $[3, \pi]$ e crescente em $[\pi, 4]$. Assim

$$\max_{x \in I} |g_1'(x)| = \max\{|g'(3)|, |g'(\pi)|, |g'(4)|\} \approx 0.346 < 1.$$

Temos também que $g_1'(x) \geq 0$ em I, logo g_1 é crescente, pelo que

$$g_1(I) = [g_1(3), g_1(4)] = [3.1411, 3.2432] \subset I.$$

Assim, pelo corolário 6.17 temos o resultado.

(e) Temos $g_2'(x) = 1 - \cos(x)$ e como para $x \in I$ temos $\cos(x) \leq 0$, concluímos que $g_2'(x) \geq 1$. O resultado sai agora do teorema de Lagrange 6.4.

(f) Ver tabelas 6.6 e 6.7.

(g) Como $f''(x) = \sin(x)$ muda de sinal em I, teremos de verificar a condição (6.15). Pela análise de monotonia de f' e f'', temos

$$C = \frac{\max_{x \in I} |f''(x)|}{2 \min_{x \in I} |f'(x)|} = \frac{|f''(4)|}{2|f'(4)|} = 0.57891$$

logo temos a condição de convergência local $|z - x_0| \leq 1/C = 1.727$, que é obviamente satisfeita para $x_0 = 3$.

(h) Ver tabelas 6.6 e 6.7.

(i) Ver tabelas 6.6 e 6.7.

(j) Neste caso temos

$$g'(z) = 1 + \cos(\pi) = 0, \quad g''(z) = -\sin(\pi) = 0, \quad g'''(z) = -\cos(\pi) \neq 0$$

logo o método do ponto fixo tem convergência cúbica. Da mesma forma, para o método de Newton temos para g da forma (6.12) que

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

e como $f''(z) = 0$, o zero de g' em z é duplo. Assim,

$$\begin{aligned} g''(z) &= \frac{[f'(z)f''(z) + f(z)f'''(z)][f'(z)]^2 - 2f(z)f'(z)[f''(z)]^2}{[f'(z)]^4} \\ &= \frac{[f'(z)f''(z) + f(z)f'''(z)]f'(z) - 2f(z)[f''(z)]^2}{[f'(z)]^3} \\ &= 0 \end{aligned}$$

e logo

$$g'''(z) = \cdots \neq 0$$

pelo que o método de Newton tem também ordem cúbica. Assim sendo, os métodos de Newton e do ponto fixo têm convergência semelhante e melhor que o método da secante neste caso.

(k) Temos a tabela 6.8, obtida a partir da expressão (6.7). Verificamos que a ordem de convergência numérica parece tender para 3 nos casos do método de Newton e do ponto fixo (como provado teoricamente), embora a partir de certa altura o logaritmo do erro entre na precisão da máquina (terceira linha) e os valores obtidos deixem de fazer sentido, pelo que foram retirados.

n	Ordem conv. numérica			
	Bisseção	Ponto Fixo	Newton	Secante
1	2.16538	3.91711	3.55789	4.40972
2	1.84476	3.234	3.15796	1.66128
3	0.751716	-	-	-

Tabela 6.8: *Exercício 6.48: Aproximação numérica* (6.7) *para a ordem de convergência dos 4 métodos.*

Capítulo 7

Sistemas Lineares de Equações

Um sistema linear de n-equações a n incógnitas $x_1, x_2, \ldots, x_n$ dado por

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

pode ser escrito na forma matricial

$$Ax = b \tag{7.1}$$

em que

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Existem duas classes de métodos para resolver o sistema matricial anterior:

- Os *métodos directos* são métodos que conduzem à solução exacta do problema, através de um número finito de operações aritméticas;
- Os métodos iterativos geram uma sucessão de aproximações da solução do problema através da aplicação sucessiva de um procedimento computacional.

A Álgebra Linear fornece alguns métodos directos para resolver o sistema anterior. Uma hipótese é inverter a matriz A e determinar a solução como

$$x = A^{-1}b.$$

Outra hipótese poderá ser eliminação de Gauss, por forma a obter uma matriz triangular superior e depois fazer substituições sucessivas. No entanto, estes métodos são apenas viáveis se a dimensão da matriz A for pequena, uma vez que noutro caso os custos computacionais envolvidos são demasiado elevados.

Nota 7.1. De facto consegue-se mostrar que para uma matriz $n \times n$, os métodos directos para resolução de sistemas envolvem cerca de $n^3/3$ somas e subtrações e $n^3/3$ multiplicações e divisões. Assim são necessárias um numero de operações total da ordem de $2n^3/3$, o que para n elevado se associa a um elevado custo computacional.

Para sistemas de dimensão elevada a melhor escolha é usar métodos iterativos, que passamos a apresentar de seguida.

7.1 Métodos iterativos para resolução de sistemas lineares

O objectivo desta secção é obter métodos iterativos para aproximar soluções de sistemas lineares em casos em que a resolução direta é demasiado morosa em termos computacionais. A base para esta abordagem é o método do ponto fixo generalizado, já exposto na secção 6.2.1.

Desta forma, pretendemos um método iterativo para resolver a equação (7.1), sem necessitar de inverter a matriz A. Uma ideia passa por escrever A na forma

$$A = M + N \tag{7.2}$$

em que M é uma matriz fácil de inverter. Assim a equação (7.1) pode ser reescrita na forma

$$Ax = b \Leftrightarrow (M+N)x = b \Leftrightarrow Mx = b - Nx \Leftrightarrow x = M^{-1}(b - Nx). \tag{7.3}$$

Considerando $Ax = M^{-1}(b - Nx)$, podemos aplicar a ideia do método do ponto fixo do teorema 6.30, obtendo

$$\begin{cases} x^{(0)} \equiv \text{iterada inicial} \\ Mx^{(k+1)} = b - Nx^{(k)} \end{cases} \quad \text{ou} \quad \begin{cases} x^{(0)} \equiv \text{iterada inicial} \\ x^{(k+1)} = M^{-1}\left(b - Nx^{(k)}\right). \end{cases} \tag{7.4}$$

em que agora a iterada $x^{(k)}$ é um vector. Note-se que em relação à notação da secção 6.2.1 passamos a iterada para o expoente, uma vez que o indice está geralmente conotado com a componente do vector no contexto de matrizes.

Note-se que a inversão de M tem de ser bastante simples, pois de outra forma não existe nenhuma vantagem neste método. De facto, diferentes escolhas da matriz M geram diferentes métodos. Vamos analisar dois casos simples.

Começamos por decompor a matriz A na soma de uma matriz triangular inferior L, uma matriz diagonal D e uma matriz triangular superior U. Assim, temos $A = L + D + U$, em que D é dada por

$$D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} \tag{7.5}$$

e as matrizes triangulares inferior L e superior U são dadas por

$$L = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}. \tag{7.6}$$

Caso escolhamos $M = D$ obtemos uma matriz de fácil inversão, uma vez que a inversa de uma matriz diagonal é a matriz diagonal cujos elementos são o inverso dos elementos da diagonal da matriz D. Obtemos assim o método de Jacobi.

Definição 7.2 (Método de Jacobi).

O método de Jacobi para a resolução do sistema

$$Ax = b$$

é dado pela escolha $M = D$ e $N = (L + U)$, o que equivale ao método iterativo

$$\begin{cases} x^{(0)} \equiv \text{iterada inicial} \\ x^{(k+1)} = D^{-1}\left(b - (L+U)x^{(k)}\right). \end{cases}$$

em que as matrizes L, D e U são dadas por (7.5) e (7.6). Componente a componente, o método pode ser dado por

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} \right). \tag{7.7}$$

Nota 7.3. Note-se que para aplicar o método de Jacobi é necessário que os elementos da diagonal sejam diferentes de zero. Se não for esse o caso, pode-se trocar linhas e/ou fazer pesquisas de pivot até termos elementos não nulos na diagonal.

Exercício 7.4. Calcule as três primeiras iterações do método de Jacobi aplicado à resolução do sistema linear

$$\begin{cases} 8x + y - 2z = 1 \\ x + 4y - z = -1 \\ x + 4z = 1 \end{cases}$$

tomando como aproximação inicial da solução $x^{(0)} = (0, 0, 0)$.

Resolução.

Na forma matricial, temos

$$\underbrace{\begin{bmatrix} 8 & 1 & -2 \\ 1 & 4 & -1 \\ 1 & 0 & 4 \end{bmatrix}}_{A} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}}_{b}$$

e tomando a notação (x_1, x_2, x_3) para as coordenadas do vector solução, de (7.7) temos

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{8}\left(1 - x_2^{(k)} + 2x_3^{(k)}\right) \\ x_2^{(k+1)} = \dfrac{1}{4}\left(-1 - x_1^{(k)} + x_3^{(k)}\right) \\ x_3^{(k+1)} = \dfrac{1}{4}\left(1 - x_1^{(k)}\right). \end{cases}$$

Assim, para $x^{(0)} = (0, 0, 0)$, obtemos

$$\begin{cases} x_1^{(1)} = \frac{1}{8}\left(1 - x_2^{(0)} + 2x_3^{(0)}\right) = 0.125 \\ x_2^{(1)} = \frac{1}{4}\left(-1 - x_1^{(0)} + x_3^{(0)}\right) = -0.25 \\ x_3^{(1)} = \frac{1}{4}\left(1 - x_1^{(0)}\right) = 0.25 \end{cases}$$

ou seja $x^{(1)} = (0.125, -0.25, 0.25)$. Aplicando de novo o método, obtemos

$$\begin{cases} x_1^{(2)} = \frac{1}{8}\left(1 - x_2^{(1)} + 2x_3^{(1)}\right) = 0.21875 \\ x_2^{(2)} = \frac{1}{4}\left(-1 - x_1^{(1)} + x_3^{(1)}\right) = -0.21875 \\ x_3^{(2)} = \frac{1}{4}\left(1 - x_1^{(1)}\right) = 0.21875 \end{cases}$$

logo a segunda iterada é dada por $x^{(2)} = (0.21875, -0.21875, 0.21875)$. A terceira iterada é então dada por

$$\begin{cases} x_1^{(3)} = \frac{1}{8}\left(1 - x_2^{(2)} + 2x_3^{(2)}\right) \approx 0.207031 \\ x_2^{(3)} = \frac{1}{4}\left(-1 - x_1^{(2)} + x_3^{(2)}\right) = -0.25 \\ x_3^{(3)} = \frac{1}{4}\left(1 - x_1^{(2)}\right) \approx 0.195313 \end{cases}$$

ou seja, $x^{(3)} = (0.207031, -0.25, 0.195313)$, o que equivale a um erro absoluto 0.003806 na norma euclidiana para vectores. Assim o método parece convergir para a solução exacta dada por $x = (0.206107, -0.251908, 0.198473)$.

Nota 7.5. Obviamente os métodos iterativos só trazem vantagem quando a dimensão n da matriz A for muito grande. De facto, o número de operações envolvidas em cada iteração do método de Jacobi é da ordem $2n^2$. Assim, em comparação com os métodos directos (ver nota 7.1), os métodos iterativos trazem vantagem se for atingida uma boa aproximação em menos de $n/3$ iterações.

Exercício Octave 7.6. Escreva uma função em Octave `Jacobi(A,b,n,x0)` que dados uma matriz `A` e um vetor `b`, determine `n` iterações do método de Jacobi para a solução do sistema $Ax = b$ com iterada inicial `x0`.

Resolução.

Temos o algoritmo no ficheiro `Jacobi.m`:

```
function [xn, lxi] = Jacobi(A,b,n,x0)
% OUTPUT: xn - iterada n;
%         lxi - lista de iteradas;
lxi=zeros(size(x0,1),n+1);
lxi(:,1)=x0;
for iter=1:n
  for ii=1:size(x0,1)
    for jj=[1:ii-1,ii+1:size(x0,1)]
      lxi(ii,iter+1)=lxi(ii,iter+1)+A(ii,jj)*lxi(jj,iter);
    end
    lxi(ii,iter+1)=(b(ii)-lxi(ii,iter+1))/A(ii,ii);
  end
end
xn=lxi(:,end);
return
```

Exercício 7.7. Verifique o resultado do exercício 7.4, recorrendo ao algoritmo criado no exercício anterior.

Outra hipótese de escolha da matriz M em (7.2) é $M = (L + D)$, que por ser uma matriz triangular inferior é também de simples inversão. Esta escolha traduz-lhe no método de Gauss-Seidel seguinte.

Definição 7.8 (Método de Gauss-Seidel).
O método de Gauss-Seidel para a resolução do sistema

$$Ax = b$$

é dado pela escolha $M = (L + D)$ e $N = U$, o que equivale a

$$\begin{cases} x^{(0)} \equiv \text{iterada inicial} \\ x^{(k+1)} = (L + D)^{-1} \left(b - Ux^{(k)}\right). \end{cases}$$

em que as matrizes L, D e U são dadas por (7.5) e (7.6). Componente a componente, o método pode ser dado por

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right). \tag{7.8}$$

Nota 7.9. Em comparação com o método de Jacobi, o método de Gauss-Seidel aproveita as componentes já actualizadas na iteração presente. Desta forma, o método de Gauss-Seidel tem (em geral) melhores resultados.

Exercício 7.10. Calcule as três primeiras iterações do método de Gauss-Seidel aplicado à resolução do mesmo sistema linear do exercício 7.4 dado por

$$\begin{cases} 8x_1 + x_2 - 2x_3 = 1 \\ x_1 + 4x_2 - x_3 = -1 \\ x_1 + 4x_3 = 1 \end{cases}$$

tomando como aproximação inicial $x^{(0)} = (0, 0, 0)$.

Resolução.

De novo, na forma matricial, temos

$$\underbrace{\begin{bmatrix} 8 & 1 & -2 \\ 1 & 4 & -1 \\ 1 & 0 & 4 \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}}_{b}$$

logo de (7.8) temos

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{8}\left(1 - x_2^{(k)} + 2x_3^{(k)}\right) \\ x_2^{(k+1)} = \dfrac{1}{4}\left(-1 - x_1^{(k+1)} + x_3^{(k)}\right) \\ x_3^{(k+1)} = \dfrac{1}{4}\left(1 - x_1^{(k+1)}\right). \end{cases}$$

Assim, para $x^{(0)} = (0, 0, 0)$, obtemos

$$\begin{cases} x_1^{(1)} = \dfrac{1}{8}\left(1 - x_2^{(0)} + 2x_3^{(0)}\right) = 0.125 \\ x_2^{(1)} = \dfrac{1}{4}\left(-1 - x_1^{(1)} + x_3^{(0)}\right) = -0.28125 \\ x_3^{(1)} = \dfrac{1}{4}\left(1 - x_1^{(1)}\right) = 0.21875 \end{cases}$$

ou seja $x^{(1)} = (0.125, -0.28125, 0.21875)$. Para a segunda iterada obtemos

$$\begin{cases} x_1^{(2)} = \frac{1}{8}\left(1 - x_2^{(1)} + 2x_3^{(1)}\right) \approx 0.214844 \\ x_2^{(2)} = \frac{1}{4}\left(-1 - x_1^{(2)} + x_3^{(1)}\right) \approx -0.249023 \\ x_3^{(2)} = \frac{1}{4}\left(1 - x_1^{(2)}\right) \approx 0.196289 \end{cases}$$

logo a segunda iterada é dada por $x^{(2)} = (0.214844, -0.249023, 0.196289)$. A terceira iterada é então dada por

$$\begin{cases} x_1^{(3)} = \frac{1}{8}\left(1 - x_2^{(2)} - 2x_3^{(2)}\right) \approx 0.2052 \\ x_2^{(3)} = \frac{1}{4}\left(-1 - x_1^{(3)} + x_3^{(2)}\right) = -0.252228 \\ x_3^{(3)} = \frac{1}{4}\left(1 - x_1^{(3)}\right) \approx 0.1987 \end{cases}$$

ou seja, $x^{(3)} = (0.2052, -0.252228, 0.1987)$. Assim o método parece convergir para a solução exacta dada por $x = (0.206107, -0.251908, 0.198473)$. De facto, o erro absoluto cometido na 3ª iterada é 0.0009876 na norma euclidiana para vectores.

Exercício Octave 7.11. Escreva uma função em Octave `GaussSeidel(A,b,n,x0)` que dados uma matriz `A` e um vetor `b`, determine `n` iterações do método de Jacobi para a solução do sistema $Ax = b$ com iterada inicial `x0`.

Resolução.

Temos o algoritmo no ficheiro `GaussSeidel.m`:

```
function [xn, lxi] = GaussSeidel(A,b,n,x0)
% OUTPUT: xn - iterada n;
%         lxi - lista de iteradas;
lxi=zeros(size(x0,1),n+1);
lxi(:,1)=x0;
for iter=1:n
  for ii=1:size(x0,1)
    for jj=1:ii-1
      lxi(ii,iter+1)=lxi(ii,iter+1)+A(ii,jj)*lxi(jj,iter+1);
    end
    for jj=ii+1:size(x0,1)
      lxi(ii,iter+1)=lxi(ii,iter+1)+A(ii,jj)*lxi(jj,iter);
    end
```

```
    lxi(ii,iter+1)=(b(ii)-lxi(ii,iter+1))/A(ii,ii);
  end
end
xn=lxi(:,end);
return
```

Exercício 7.12. Verifique o resultado do exercício 7.10, recorrendo ao algoritmo criado no exercício anterior.

Como em todos os métodos numéricos, é importante ter critérios para saber quando o método converge ou diverge. Para isso, tomemos o método geral (7.4) tal que

$$x^{(k+1)} = M^{-1}\left(b - Nx^{(k)}\right)$$

que equivale a iterações do método do ponto fixo generalizado, com $Ax =$

Desta forma, o erro na iterada $(k+1)$ é dado por

$$e^{(k+1)} = \left|\left|x - x^{(k+1)}\right|\right| = \left|\left|x - M^{-1}\left(b - Nx^{(k)}\right)\right|\right|.$$

Note-se agora que de (7.3) temos para a solução exacta x

$$x = M^{-1}(b - Nx)$$

pelo que obtemos

$$\begin{aligned} e^{(k+1)} &= \left|\left|M^{-1}(b - Nx) - M^{-1}\left(b - Nx^{(k)}\right)\right|\right| \\ &= \left|\left|-M^{-1}N\left(x - x^{(k)}\right)\right|\right| \\ &\leq \left|\left|-M^{-1}N\right|\right| \left|\left|x - x^{(k)}\right|\right| \\ &= \left|\left|-M^{-1}N\right|\right| e^{(k)}. \end{aligned}$$

Repetindo o processo obtemos

$$e^{(k+1)} \leq ||C||^{k+1}\, e^{(0)}. \tag{7.9}$$

para a matriz $C = -M^{-1}N$, que se denomina por **matriz do método iterativo** (7.4). Dito de outra forma no contexto do método do ponto fixo generalizado da secção 6.2.1, temos que o operador é uma contração se e só se $\|C\| = L < 1$.

Recordando as definições de raio espectral de uma matriz (4.9) e a sua propriedade (4.12) temos que se $\rho(C) < 1$ então existe uma norma $||.||$ tal que $||C|| < 1$ e logo o método converge nessa norma. Por outro lado, de (4.11) temos que

se $\rho(C) > 1$ então $||C|| > 1$ para qualquer norma, logo existe um $e^{(0)}$ (o vector próprio associado ao maior valor próprio) tal que

$$e^{(k+1)} = ||C||^{k+1} e^{(0)} \to \infty, \quad k \to \infty,$$

pelo que o método é divergente para a iterada inicial $x^{(0)}$ correspondente. Temos então o seguinte teorema, que pode ser visto como um corolário do teorema do ponto fixo generalizado 6.30.

Teorema 7.13 (Condição necessária e suficiente de convergência).
O método da forma (7.4) é convergente para qualquer iterada inicial $x^{(0)}$ ***se e só se*** *$\rho(C) < 1$, para $C = -M^{-1}N$.*

Deste resultado geral podemos tirar condições necessárias e suficientes de convergência para os métodos de Jacobi e Gauss-Seidel.

Corolário 7.14 (Condição necessária e suficiente de convergência para Jacobi).
O método de Jacobi (7.7) é convergente para qualquer iterada inicial $x^{(0)}$ ***se e só se*** *$\rho(C) < 1$, para $C = -D^{-1}(L + U)$ a matriz do método iterativo de Jacobi.*

Corolário 7.15 (Condição necessária e suficiente de convergência para Gauss-Seidel).
O método de Gauss-Seidel (7.8) é convergente para qualquer iterada inicial $x^{(0)}$ ***se e só se*** *$\rho(C) < 1$, para $C = -(L + D)^{-1}U$ a matriz do método iterativo de Gauss-Seidel.*

A condição anterior é de difícil verificação na prática, ainda para mais em sistemas de elevada dimensão para os quais os métodos iterativos são adequados. Mais ainda, esta condição exige o cálculo de uma matriz C a partir da decomposição da matriz A. Por estas razões é necessário criar outras condições que garantam convergência mas que sejam de mais fácil verificação. Temos então a seguinte definição.

Definição 7.16 (Matriz com diagonal estritamente dominante por linhas ou colunas).
Diz-se que a matriz quadrada A de dimensão n tem **diagonal estritamente dominante por linhas** se

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,$$

para qualquer $i = 1, 2, \ldots, n$, isto é, se o módulo do elemento da diagonal for maior que a soma dos módulos dos restantes elementos da linha.

De forma semelhante, diz-se que a matriz quadrada A de dimensão n tem **diagonal estritamente dominante por colunas** se

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|,$$

para qualquer $j = 1, 2, \ldots, n$, isto é, se o módulo do elemento da diagonal for maior que a soma dos módulos dos restantes elementos da coluna.

Temos então a seguinte condição de convergência.

Teorema 7.17 (Condição suficiente de convergência).
Se a matriz A do sistema

$$Ax = b$$

tem diagonal estritamente dominante por linhas ou por colunas, então os métodos de Jacobi (7.7) e de Gauss-Seidel (7.8) são convergentes para qualquer iterada inicial $x^{(0)}$.

Demonstração. Faremos a prova apenas no caso do método de Jacobi, por simplicidade. O caso do Método de Gauss-Seidel pode ser encontrado em [5, 10]. Por definição, a matriz $C = -D^{-1}(L + U)$ tem entradas

$$c_{ij} = \begin{cases} 0, & i = j \\ -a_{ij}/a_{ii}, & i = j \end{cases}$$

em que a_{ij} são as entradas da matriz A do sistema linear, uma vez que a inversa de uma matriz diagonal é a matriz diagonal com o inverso das entradas. Temos que a norma das linhas de C é dada por

$$\|C\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |c_{ij}| = \max_{i=1,\ldots,n} \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|}.$$

logo supondo que A é estritamente dominante por linhas, temos

$$\sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

logo tomando o máximo L dos valores anteriores para $i = 1, \ldots, n$, obtemos

$$\|C\|_\infty = L < 1.$$

A propriedade (4.11) e o teorema 7.14 terminam a prova neste caso. Caso a matriz seja estritamente dominante por colunas, a prova sai considerando a norma das colunas. □

Nota 7.18. Ao contrário do teorema 7.13, o resultado 7.17 dá-nos apenas uma condição suficiente de convergência. Desta forma, se a matriz A não for estritamente dominante por linhas ou colunas não quer dizer que os métodos de Jacobi ou Gauss-Seidel não convirjam.

Exercício 7.19. Indique se pode garantir pelo teorema 7.17 que os métodos de Jacobi e Gauss-Seidel são convergentes para os sistemas dados, e nesses casos determine as 3 primeiras iteradas de cada um deles.

(a) $\begin{cases} 2x + y = 1 \\ 2x + 3y = -3 \end{cases}$

(b) $\begin{cases} x + y - 4z = 1 \\ x + 4y + z = -2 \\ 3x + y + z = 2 \end{cases}$

(c) $\begin{cases} 2x + y - z = 1 \\ x - 3y + z = -2 \\ x + y + 3z = 2 \end{cases}$

(d) $\begin{bmatrix} 4 & 10 & -1 \\ 7 & 7 & -1 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$

Resposta.

Em cada uma das alíneas consideramos $x^{(0)} = \vec{0}$, ainda que qualquer outro valor seja admissível.

(a) Na forma matricial

$$Ax = b$$

temos

$$A = \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -3 \end{bmatrix}.$$

A matriz A tem diagonal estritamente dominante por linhas, logo tanto o método de Jacobi como o de Gauss-Seidel convergem para qualquer iterada inicial.

Assim pelo método de Jacobi, temos

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(1 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{3}\left(-3 - 2x_1^{(k)}\right) \end{cases}$$

pelo que obtemos

$$\begin{cases} x^{(0)} = (0, 0) \\ x^{(1)} = (0.5, -1) \\ x^{(2)} = (1, -1.33333) \\ x^{(3)} = (1.16667, -1.66667) \end{cases}$$

que converge para a solução exacta $x = (1.5, -2)$.

Pelo método de Gauss-Seidel, temos

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(1 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{3}\left(-3 - 2x_1^{(k+1)}\right) \end{cases}$$

pelo que obtemos

$$\begin{cases} x^{(0)} = (0, 0) \\ x^{(1)} = (0.5, -1.33333) \\ x^{(2)} = (1.16667, -1.77778) \\ x^{(3)} = (1.38889, -1.92593)\,. \end{cases}$$

(b) Temos o sistema matricial correspondente

$$Ax = b$$

com

$$A = \begin{bmatrix} 1 & 1 & -4 \\ 1 & 4 & 1 \\ 3 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}.$$

A matriz A não tem diagonal estritamente dominante por linhas, nem por colunas, logo nada se pode concluir sobre a convergência do método. No

entanto se trocarmos as 1ª e 3ª linhas de A ficamos com uma matriz com diagonal estritamente dominante por linhas. Assim, fazendo essa troca e a troca correspondente no vector b, obtemos

$$A = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & -4 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}.$$

Assim já podemos aplicar os métodos de Jacobi e Gauss-Seidel com garantia de convergência. Para o método de Jacobi

$$\begin{cases} x_1^{(k+1)} = \frac{1}{3}\left(2 - x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{4}\left(-2 - x_1^{(k)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = -\frac{1}{4}\left(1 - x_1^{(k)} - x_2^{(k)}\right) \end{cases}$$

obtemos

$$\begin{cases} x^{(0)} = (0, 0, 0) \\ x^{(1)} = (0.666667, -0.5, -0.25) \\ x^{(2)} = (0.916667, -0.604167, -0.208333) \\ x^{(3)} = (0.937500, -0.677083, -0.171875), \end{cases}$$

enquanto que para o método de Gauss-Seidel

$$\begin{cases} x_1^{(k+1)} = \frac{1}{3}\left(2 - x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{4}\left(-2 - x_1^{(k+1)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = -\frac{1}{4}\left(1 - x_1^{(k+1)} - x_2^{(k+1)}\right) \end{cases}$$

obtemos

$$\begin{cases} x^{(0)} = (0, 0, 0) \\ x^{(1)} = (0.666667, -0.666667, -0.25) \\ x^{(2)} = (0.972222, -0.680556, -0.177084) \\ x^{(3)} = (0.952547, -0.693866, -0.185330). \end{cases}$$

que convergem para a solução $x \approx (0.959184, -0.693878, -0.183673)$.

(c) A matriz A correspondente é

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

que não tem diagonal estritamente dominante por colunas nem por linhas. Neste caso, trocas de linhas e colunas não iriam solucionar o problema, pelo que não podemos garantir convergência pela condição suficiente do teorema 7.17. No entanto, para as matrizes C associadas ao método de Jacobi e Gauss-Seidel, obtemos os raios espectrais $\rho(C) = 0.666667 < 1$ e $\rho(C) = 0.593592 < 1$, respectivamente, pelo que o método é convergente para a solução $x \approx (0.538462, 0.807692, -0.115385)$.

(d) Trocando a primeira com a segunda linha, obtemos

$$\begin{bmatrix} 7 & 7 & -1 \\ 4 & 10 & -1 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix},$$

logo a matriz tem diagonal estritamente dominante por colunas, pelo que os métodos de Jacobi e Gauss-Seidel convergem para a solução exacta

$$x = (0.342282, -0.221477, -0.154362).$$

Assim obtemos para o método de Jacobi

$$x^{(3)} = (0.291156, -0.190952, -0.109524)$$

e para o método de Gauss-Seidel

$$x^{(3)} = (0.338290, -0.222833, -0.151249).$$

Exercício 7.20. Determine os erros absolutos cometidos na norma euclidiana para vectores pela 3ª iteração dos métodos de Jacobi e Gauss-Seidel no exercício anterior.

Resolução.

(a) Jacobi: $e^{(3)} = 0.4714$; Gauss-Seidel: $e^{(3)} = 0.133536$;

(b) Jacobi: $e^{(3)} = 0.0282021$; Gauss-Seidel: $e^{(3)} = 0.00566184$;

(c) Jacobi: $e^{(3)} = 0.292880$; Gauss-Seidel: $e^{(3)} = 0.167038$;

(d) Jacobi: $e^{(3)} = 0.0745388$; Gauss-Seidel: $e^{(3)} = 0.00524143$;

Capítulo 8

Integração numérica

Nesta secção vamos estudar formas de aproximar integrais definidos de funções conhecidas em apenas alguns pontos ou de funções das quais não se conhecem formas analíticas de primitivação. Tal como nos integrais, a integração numérica tem aplicações no cálculo de áreas, na resolução de equações integrais e na reconstrução de funções a partir das suas derivadas, entre outros.

Nota 8.1. Em Octave, pode utilizar o comando `quad(f,a,b)` em que a função integranda é definida por `f=inline(expf)` em que `expf` é a expressão de f para obter uma aproximação (numérica) do valor do integral $\int_a^b f(x)dx$. Esta aproximação é obtida por processos numéricos adaptativos, que procuram obter uma aproximação com a precisão da máquina (ou seja, 10^{-16}) baseando-se em regras de quadratura como as que vamos descrever neste capítulo.

A ideia passa por considerar $(n+1)$ pontos

$$a = x_0 < x_1 < \cdots < x_n = b$$

no intervalo $[a, b]$ de integração e considerar uma aproximação da forma

$$\int_a^b f(x)dx \approx Q_n(f)$$

em que $Q_n(f)$ é chamada uma **regra de quadratura** ou **fórmula de integração numérica** e é da forma

$$Q_n(f) = \sum_{j=0}^{n} \alpha_j \, f(x_j). \tag{8.1}$$

Aos pontos x_j, $j = 0, 1, \ldots n$, chamam-se **nós de quadratura** e aos valores reais α_j chamam-se os **pesos da quadratura**.

Definição 8.2 (Regra de Quadratura exata).

Diz-se que uma regra de quadratura é **exata** para a função f, se

$$\int_a^b f(x)\,dx = Q_n(f).$$

Assim, diz-se que uma regra de quadratura é **exata para polinómios de grau menor ou igual a n** se

$$\int_a^b p(x)\,dx = Q_n(p)$$

para qualquer polinómio p de grau menor ou igual a n.

Uma forma de obter regras de quadratura é considerar a aproximação

$$\int_a^b f(x)\,dx \approx \int_a^b p_n(x)\,dx$$

em que $p_n = p_n(x)$ é um polinómio interpolador de f de ordem n. Diferentes escolhas de nós de interpolação e de ordem do polinómio levam a diferentes regras de quadratura. Nas linhas que se seguem vamos estudar algumas regras de quadratura, começando por considerar nós igualmente espaçados.

8.1 Quadraturas simples com nós igualmente espaçados

Nesta secção vamos considerar que os nós de quadratura estão igualmente espaçados, isto é, o espaçamento entre os nós é sempre igual. Assim, considerando um espaçamento $h = (b-a)/n$, temos a relação

$$x_{j+1} = x_j + h, \quad j = 0, 1, \ldots n-1,$$

ou seja,

$$x_0 = a, \quad x_1 = a + h, \quad x_2 = a + 2h, \quad \ldots, x_n = a + nh (= b).$$

Neste caso podemos considerar as **regras de Newton-Cotes** de ordem n que passam por considerar a aproximação do integral

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)dx$$

em que p_n é o polinómio interpolador de grau n em nós de quadratura igualmente espaçados. Começamos por considerar regras simples, em que a aproximação é feita globalmente em todo o intervalo. No entanto, devido ao fenómeno de Runge, quando o grau do polinómio aumenta a aproximação do integral pode deteriorar-se. Nesse caso é preferível considerar regras compostas, em que se considera uma aproximação seccionalmente polinomial, como veremos na secção 8.2.

8.1.1 Regra dos trapézios simples

Começamos pelo exemplo mais simples, isto é, consideramos a aproximação

$$\int_a^b f(x)dx \approx \int_a^b p_1(x)dx$$

em que p_1 é o polinómio interpolador de f nos nós

$$x_0 = a, \quad x_1 = b,$$

e

$$f_0 = f(a), \quad f_1 = f(b),$$

isto é, consideramos um espaçamento $h = b - a$. Na forma de Lagrange o polinómio interpolador é dado por

$$p_1(x) = f_0 \frac{x - x_1}{x_0 - x_1} + f_1 \frac{x - x_0}{x_1 - x_0} = \frac{(f_1 - f_0)x + f_0 b - f_1 a}{b - a},$$

logo obtemos a regra de quadratura

$$\begin{aligned} Q(f) &= \int_a^b p_1(x)\,dx \\ &= \frac{1}{b-a}\int_a^b (f_1-f_0)x + f_0 b - f_1 a dx \\ &= \frac{1}{b-a}\left[(f_1-f_0)\frac{x^2}{2} + (f_0 b - f_1 a)x\right]_a^b \\ &= \frac{1}{b-a}\left\{(f_1-f_0)\left(\frac{b^2}{2}-\frac{a^2}{2}\right) + (f_0 b - f_1 a)(b-a)\right\} \\ &= \frac{1}{b-a}\left\{\frac{f_1}{2}(b-a)^2 + \frac{f_0}{2}(b-a)^2\right\} \\ &= h\left(\frac{f_0+f_1}{2}\right). \end{aligned}$$

Outra forma de obter este resultado é através da área de um trapézio. Como o integral definido de f representa a área compreendida entre o gráfico de f e o eixo dos xx, a área sob o gráfico do polinómio interpolador p_1 define um trapézio (ver figura 8.1) de área

$$\int_a^b p_1(x)dx = \frac{(f(b)+f(a))(b-a)}{2} = h\,\frac{(f_1+f_0)}{2}.$$

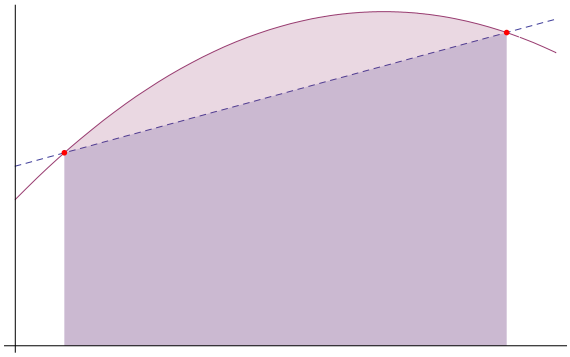


Figura 8.1: *Área sob os gráficos de f (linha contínua) e do seu polinómio interpolador de grau 1 (a tracejado).*

Definimos assim a regra dos trapézios simples.

Definição 8.3 (Regra dos trapézios simples).

Seja $f : [a, b] \to \mathbb{R}$ uma função contínua em $[a, b]$ e seja $h = (b - a)$ o comprimento do intervalo.

Então, o integral definido de f em $[a, b]$ pode ser aproximado pela regra dos trapézios

$$\int_a^b f(x)\,dx \approx Q^T(f)$$

em que

$$Q^T(f) = h\left(\frac{f(a) + f(b)}{2}\right). \tag{8.2}$$

Nota 8.4. A regra dos trapézios é uma regra de quadratura da forma (8.1), com nós de quadratura

$$x_0 = a, \quad x_1 = b,$$

e pesos de quadratura

$$\alpha_0 = \alpha_1 = \frac{h}{2}.$$

Exercício 8.5. Aproxime os integrais seguintes pela regra dos trapézios.

(a) $\int_0^1 e^x\,dx$;

(b) $\int_0^\pi \sin x\,dx$;

(c) $\int_0^\pi \cos x\,dx$;

(d) $\int_{-2}^2 \frac{x-1}{2}\,dx$;

Resposta.

(a) $Q^T(f) = \frac{e+1}{2} \approx 1.85914$, enquanto que o valor exato é $\int_0^1 f(x)\,dx = e - 1 \approx 1.71828$;

(b) $Q^T(f) = 0$, enquanto que o valor exato é $\int_0^\pi f(x)\,dx = 2$;

(c) $Q^T(f) = 0$, que por coincidência coincide com o valor exato $\int_0^\pi f(x)\,dx = 0$;

(d) $Q^T(f) = -2$, que coincide com o valor exato $\int_0^\pi f(x)\,dx = -2$.;

Após resolver as duas últimas alíneas do exercício anterior, verificamos que convém clarificar as funções para as quais a regra dos trapézios é exata. Temos o seguinte resultado.

Teorema 8.6.

A regra dos trapézios simples (8.2) é exata para polinómios de grau menor ou igual a 1.

Demonstração. Vem diretamente da definição, uma vez que o polinómio interpolador de grau 1 de um polinómio de primeiro grau coincide com esse mesmo polinómio. □

Pretendemos também estabelecer uma fórmula de erro para a regra de quadratura anterior. Para isso vamos precisar do resultado seguinte.

Teorema 8.7 (Teorema do Valor Médio para Integrais).

Sejam $f, g : [a, b] \to \mathbb{R}$ *contínuas em* $[a, b]$*, tais que* g *não muda de sinal em* $[a, b]$.

Então existe um $\zeta \in [a, b]$ *tal que*

$$\int_a^b f(x)\, g(x)\, dx = f(\zeta) \int_a^b g(x)\, dx.$$

Demonstração. Este é um resultado base para a teoria de integração e a prova pode ser encontrada em qualquer livro de análise integral, como por exemplo [7, 8, 12]. □

Temos então o seguinte teorema. Para simplificar a notação vamos utilizar

$$I(f) := \int_a^b f(x)dx.$$

Teorema 8.8 (Erro da regra dos trapézios).

Seja $f : [a, b] \to \mathbb{R}$ *uma função duas vezes diferenciável e seja* $h = (b - a)$ *o comprimento do intervalo.*

Então o erro da regra dos trapézios é dado por

$$I(f) - Q^T(f) = -\frac{h^3}{12} f''(\xi),$$

para algum $\xi \in [a, b]$*. Desta forma temos o majorante para o erro absoluto*

$$e^T(f) = \left| I(f) - Q^T(f) \right| \leq \frac{h^3}{12} \max_{x \in [a,b]} \left| f''(x) \right|. \tag{8.3}$$

Demonstração. Temos a fórmula do erro para o polinómio interpolador de grau 1 (5.8) dada por

$$f(x) - p_1(x) = f[x_0, x_1, x](x - x_0)(x - x_1) = f[a, b, x](x - a)(x - b).$$

Como $(x - a) \geq 0$ e $(x - b) \leq 0$ para $x \in [a, b]$ temos $(x - a)(x - b) \leq 0$. Assim sendo, temos

$$\begin{aligned} I(f) - Q^T(f) &= \int_a^b f(x)\,dx - \int_a^b p_1(x)\,dx \\ &= \int_a^b (f(x) - p_1(x))\,dx \\ &= \int_a^b f[a, b, x](x - a)(x - b)\,dx \\ &= f[a, b, \zeta] \int_a^b (x - a)(x - b)\,dx \\ &= f[a, b, \zeta] \int_a^b x^2 - (a + b)x + ab\,dx \\ &= -f[a, b, \zeta]\frac{(b - a)^3}{6} \end{aligned}$$

para algum $\zeta \in [a, b]$, por aplicação do Teorema do Valor Médio para integrais. Aplicando agora o teorema 5.28 temos que existe um $\xi \in [a, b]$, tal que

$$f[a, b, \zeta] = \frac{f''(\xi)}{2}$$

de onde sai diretamente o resultado. □

Nota 8.9. Mais uma vez é claro que se a função $f = f(x)$ for um polinómio de grau 1, então $f''(x) = 0$ e logo o erro é nulo, pelo que a regra dos trapézios é exata para polinómios de grau 1.

Exercício 8.10. Determine as aproximações dos integrais seguintes pela regra dos trapézios, determinando, um majorante para o erro absoluto, se possível.

(a) $\int_0^2 e^x\,dx$;

(b) $\int_0^{\frac{\pi}{4}} \sin x\,dx$;

(c) $\int_0^{\frac{\pi}{3}} x \cos x\,dx$;

(d) $\int_1^2 \ln(x)\,dx$;

(e) $\int_0^1 \sqrt{x}\,dx$;

Resposta.

(a) $Q^T(f) = e^2 + 1 \approx 8.38906$, e temos a estimativa de erro

$$e^T(f) \leq \frac{2^3}{12} \max_{x\in[0,2]} |e^x| = \frac{2^3 e^2}{12} \approx 4.92604.$$

O valor exato é $\int_0^2 f(x)\,dx = e^2 - 1 \approx 6.38906$;

(b) $Q^T(f) = \frac{\pi\sqrt{2}}{16} \approx 0.27768$, e temos a estimativa de erro

$$e^T(f) \leq \frac{\pi^3}{12 \times 4^3} \max_{x\in[0,\pi/4]} |\sin(x)| = \frac{\pi^3\sqrt{2}}{24 \times 4^3} \approx 0.0285478.$$

O valor exato é $\int_0^{\frac{\pi}{4}} f(x)\,dx = 1 - \frac{\sqrt{2}}{2} \approx 0.292893$;

(c) $Q^T(f) = \frac{\pi^2}{36} \approx 0.274156$, e temos a estimativa de erro

$$\begin{aligned} e^T(f) &\leq \frac{\pi^3}{12 \times 3^3} \max_{x\in[0,\pi/3]} |-2\sin(x) - x\cos x| \\ &\leq \frac{\pi^3}{12 \times 3^3} \left(2 + \max_{x\in[0,\pi/3]} |x|\right) \\ &= \frac{\pi^3}{12 \times 3^3} \left(2 + \frac{\pi}{3}\right) \\ &\approx 0.291612. \end{aligned}$$

O valor exato é $\int_0^{\frac{\pi}{3}} f(x)\,dx = \frac{1}{6}\left(-3 + \sqrt{3}\pi\right) \approx 0.4069$;

(d) $Q^T(f) = \frac{\ln(2)}{2} \approx 0.346574$, e temos a estimativa de erro

$$e^T(f) \leq \frac{1^3}{12} \max_{x\in[1,2]} \left|x^{-2}\right| = \frac{1}{12} \approx 0.0833333.$$

O valor exato é $\int_1^2 f(x)\,dx = -1 + \ln(4) \approx 0.386294$.

(e) $Q^T(f) = \frac{1}{2} = 0.5$, mas não podemos aplicar a estimativa de erro pois a função $f(x) = \sqrt{x}$ não é diferenciável em $x = 0$. O valor exato é $\int_0^1 f(x)\,dx = \frac{2}{3} \approx 0.66666$.

8.1.2 Regra de Simpson simples

Vimos nos exercícios anteriores que as aproximações pela regra dos trapézios são, em geral, bastante grosseiras. Desta forma podemos considerar uma aproximação para o integral utilizando um polinómio interpolador de grau 2, isto é,

$$\int_a^b f(x)\,dx \approx \int_a^b p_2(x)\,dx$$

em que p_2 é o polinómio interpolador de f de grau 2 nos nós igualmente espaçados

$$x_0 = a, \quad x_1 = \frac{b+a}{2}, \quad x_2 = b,$$

com espaçamento $h = \frac{b+a}{2}$.

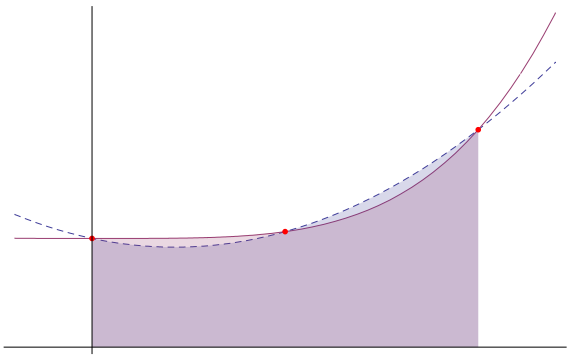


Figura 8.2: *Área sob os gráficos de f (linha contínua) e do seu polinómio interpolador de grau 2 (a tracejado).*

A regra de quadratura que se obtém é chamada regra de Simpson e é dada pelo seguinte teorema.

Definição 8.11 (Regra de Simpson simples).
Seja $f : [a, b] \to \mathbb{R}$ uma função contínua em $[a, b]$ e seja $h = (b - a)/2$ o espaçamento entre os nós de quadratura considerados

$$x_0 = a, \quad x_1 = \frac{b+a}{2}, \quad x_2 = b.$$

Então, o integral definido de f em $[a, b]$ pode ser aproximado pela regra de Simpson

$$\int_a^b f(x)\,dx \approx Q^S(f)$$

em que

$$Q^S(f) = h\left(\frac{f(x_0) + 4f(x_1) + f(x_2)}{3}\right). \tag{8.4}$$

Nota 8.12. A regra de Simpson é uma regra de quadratura da forma (8.1), com nós de quadratura

$$x_0 = a, \quad x_1 = \frac{b+a}{2}, \quad x_2 = b,$$

e pesos de quadratura

$$\alpha_0 = \frac{h}{3}, \quad \alpha_1 = \frac{4h}{3}, \quad \alpha_2 = \frac{h}{3}.$$

Como a regra de Simpson é obtida por integração exata do polinómio interpolador de grau 2, temos o seguinte resultado.

Teorema 8.13.
A regra de Simpson simples (8.2) é exata para polinómios de grau menor ou igual a 2.

Exercício 8.14. Aplique a regra de Simpson ao mesmos integrais do exercício 8.5, isto é, a

(a) $\int_0^1 e^x\,dx$;

(b) $\int_0^\pi \sin x\,dx$;

(c) $\int_0^\pi \cos x\,dx$;

(d) $\int_{-2}^2 \frac{x-1}{2}\,dx$;

Resposta.

(a) $Q^S(f) = \frac{e+4\sqrt{e}+1}{6} \approx 1.71886$, enquanto que o valor exato é $\int_0^1 f(x)\,dx = e - 1 \approx 1.71828$;

(b) $Q^S(f) = 2.0944$, enquanto que o valor exato é $\int_0^\pi f(x)\,dx = 2$;

(c) $Q^S(f) = 0$, que coincide com o valor exato $\int_0^\pi f(x)\,dx = 0$;

(d) $Q^S(f) = -2$, que coincide com o valor exato $\int_0^\pi f(x)\,dx = -2$.;

Pelo exemplo anterior verificamos que a aproximação pelo polinómio de Simpson é em geral melhor que pela regra dos trapézios. Para funções 3-vezes diferenciáveis isto é esperado, pois o erro do polinómio interpolador de grau 2 é também em geral menor que o do polinómio interpolador de grau 1. Matematicamente, este fenómeno pode ser explicado pela fórmula do erro da fórmula de Simpson.

Teorema 8.15 (Erro da regra de Simpson).
Seja $f : [a, b] \to \mathbb{R}$ uma função quatro vezes diferenciável e seja $h = (b - a)/2$ o espaçamento entre os nós de quadratura

$$x_0 = a, \quad x_1 = \frac{b+a}{2}, \quad x_2 = b.$$

Então o erro da regra de Simpson é dado por

$$I(f) - Q^S(f) = -\frac{h^5}{90} f^{(4)}(\xi), \tag{8.5}$$

para algum $\xi \in [a, b]$. Desta forma temos o majorante para o erro absoluto

$$e^S(f) = \left| I(f) - Q^S(f) \right| \leq \frac{h^5}{90} \max_{x \in [a,b]} \left| f^{(4)}(x) \right|. \tag{8.6}$$

Demonstração. A demonstração é semelhante à regra dos trapézios (ver teorema 8.8). Tomando o erro do polinómio interpolador de grau 2 (ver (5.8)) dado por

$$f(x) - p_2(x) = f[x_0, x_1, x_2, x](x - x_0)(x - x_1)(x - x_2),$$

a dificuldade passa agora pelo facto da expressão $(x - x_0)(x - x_1)(x - x_2)$ oscilar de sinal para $x \in [a, b]$ e logo não se poder aplicar diretamente o teorema do valor médio para integrais 8.7. Para ultrapassar esta dificuldade, adicionamos um novo nó $x_3 \in [a, b]$ e obtemos

$$f[x_0, x_1, x_2, x_3, x] = \frac{f[x_0, x_1, x_2, x_3] - f[x_0, x_1, x_2, x]}{x - x_3}$$

o que equivale a

$$f[x_0, x_1, x_2, x] = -f[x_0, x_1, x_2, x_3, x](x - x_3) + f[x_0, x_1, x_2, x_3],$$

uma vez que podemos alterar a ordem dos pontos das diferenças divididas. Tomando a expressão anterior e fazendo $x_3 \to x_1$, temos o resultado por aplicação do teorema do valor médio para integrais e do teorema 5.28. □

Nota 8.16. O teorema anterior permite-nos melhorar o resultado do teorema 8.13. Na realidade o erro da regra de Simpson é nulo se a quarta derivada de f for nula, ou seja, a regra de Simpson é exata para polinómios de grau menor ou igual a 3.

Em comparação com o erro da regra dos trapézios (8.3), o majorante do erro da regra de Simpson (8.6) apresenta um denominador maior e um expoente do espaçamento h também maior, o que prevê melhores resultados para espaçamentos h pequenos. No entanto é necessário comparar o comportamento da segunda e quarta derivadas da função f, para determinar qual a melhor aproximação.

Exercício 8.17. Determine as aproximações dos integrais do exercício 8.10, determinando se possível um majorante para o erro absoluto.

(a) $\int_0^2 e^x\,dx$;

(b) $\int_0^{\frac{\pi}{4}} \sin x\,dx$;

(c) $\int_0^{\frac{\pi}{3}} x\cos x\,dx$;

(d) $\int_1^2 \ln(x)\,dx$;

(e) $\int_0^1 \sqrt{x}\,dx$;

Resposta.

(a) $Q^S(f) = \frac{e^2+4e+1}{3} \approx 6.42073$ e temos o majorante do erro

$$e^S(f) \leq \frac{1^5}{90} \max_{x\in[0,2]} |e^x| = \frac{e^2}{90} \approx 0.0821006.$$

O valor exato é $\int_0^2 f(x)\,dx = e^2 - 1 \approx 6.38906$;

(b) $Q^S(f) = 0.292933$ e temos a estimativa de erro

$$e^S(f) \leq \frac{\pi^5}{90\times 8^5} \max_{x\in[0,\pi/4]} |\sin(x)| = \frac{\pi^5\sqrt{2}}{180\times 8^5} \approx 0.000073374.$$

O valor exato é $\int_0^{\frac{\pi}{4}} f(x)\,dx = 1 - \frac{\sqrt{2}}{2} \approx 0.292893$;

(c) $Q^S(f) = 0.407953$ e temos a estimativa de erro

$$\begin{aligned} e^S(f) &\leq \frac{\pi^5}{90 \times 6^5} \max_{x\in[0,\pi/3]} |4\sin(x) + x\cos x| \\ &\leq \frac{\pi^5}{90 \times 6^5} \left(4 + \max_{x\in[0,\pi/3]} |x|\right) \\ &= \frac{\pi^5}{90 \times 6^5} \left(4 + \frac{\pi}{3}\right) \\ &\approx 0.00220699. \end{aligned}$$

O valor exato é $\int_0^{\frac{\pi}{3}} f(x)\,dx = \frac{1}{6}\left(-3 + \sqrt{3}\pi\right) \approx 0.4069$;

(d) $Q^S(f) = 0.385835$, e temos a estimativa de erro

$$e^S(f) \leq \frac{0.5^5}{90} \max_{x\in[1,2]} \left|6x^{-4}\right| = \frac{1}{480} \approx 0.00208333.$$

O valor exato é $\int_1^2 f(x)\,dx = -1 + \ln(4) \approx 0.386294$.

(e) $Q^S(f) = \frac{1}{2} = 0.638071$, mas não podemos aplicar a estimativa de erro pois a função $f(x) = \sqrt{x}$ não é diferenciável em $x = 0$.
O valor exato é $\int_0^1 f(x)\,dx = \frac{2}{3} \approx 0.66666$.

8.1.3 Regras de Newton-Cotes de ordem superior

Uma possibilidade para encontrar regras de quadratura com melhores aproximações é aumentar o grau n do polinómio interpolador p_n de f nos nós igualmente espaçados

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + n\,h = b,$$

com espaçamento $h = \frac{b-a}{n}$. Desta forma obtemos a regra de quadratura

$$Q_n(f) = \int_a^b p_n(x)\,dx$$

e a consequente aproximação

$$\int_a^b f(x)\,dx = Q_n(f).$$

Como exemplo deixamos a regra dos 3/8 (três-oitavos) exata para polinómios de grau $n = 3$ dada por

$$Q^{3/8}(f) = \frac{3}{8}h\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right], \tag{8.7}$$

com erro

$$I(f) - Q^{3/8}(f) = -\frac{3h^5}{80}f^{(4)}(\xi) \tag{8.8}$$

para algum $\xi \in [a, b]$. Tomando em consideração a nota 8.16, a regra dos 3/8 é tão exata como a de Simpson, uma vez que ambas são exatas para polinómios de grau 3. Se tomarmos $n = 4$ obtemos a regra de Milne dada por

$$Q^{M}(f) = \frac{2}{45}h\left[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\right], \tag{8.9}$$

com erro

$$I(f) - Q^{M}(f) = -\frac{8h^7}{945}f^{(6)}(\xi) \tag{8.10}$$

para algum $\xi \in [a, b]$. Chamamos a atenção que a regra de Milne é exata para polinómios de grau 5 (pois nesse caso a derivada de ordem seis anula-se), ganhando um grau de exatidão ao esperado, à semelhança do que já tinha acontecido para a regra de Simpson.

Note-se que à medida que o grau do polinómio considerado aumenta, o processo de obter os pesos de integração fica cada vez mais complexo (voltaremos a este assunto, quando estudarmos regras de quadratura para nós com espaçamento variável). Além disso, o fenómeno de Runge verificado para interpolação polinomial de grau elevado com nós igualmente espaçados pode deteriorar muito a aproximação. Assim, uma possibilidade é considerar regras de quadratura compostas.

8.2 Quadraturas compostas com nós igualmente espaçados

A ideia das regras composta é conceptualmente bastante simples e passa por dividir o intervalo $[a, b]$ em n sub-intervalos e aplicar uma regra simples em cada um dos sub-intervalos. Assim, consegue-se acompanhar a curvatura da função integranda f sem aumentar o grau do polinómio interpolador, ou seja, sem se verificar o fenómeno de Runge. Nesta secção vamos considerar os nós igualmente espaçados

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + n\,h = b,$$

com espaçamento $h = \frac{b-a}{n}$ e $f_0, f_1, \ldots, f_n$ os valores da função f nos respetivos nós de quadratura. Começamos por ilustrar para o caso mais simples, a regra dos trapézios composta.

8.2.1 Regra dos trapézios composta

No contexto dos nós anteriores, podemos aplicar a regra dos trapézios simples em cada um dos sub-intervalos $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$. Isto é equivalente a considerar o polinómio interpolador de grau 1, nos extremos de cada sub-intervalo.

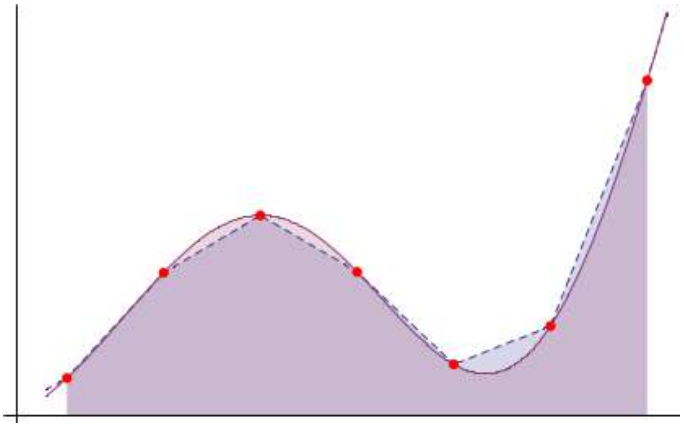


Figura 8.3: *Área sob os gráficos de f (linha contínua) e do polinómio interpolador de grau 1 em cada subintervalo (a tracejado).*

A aproximação

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,dx \\ &\approx h\left(\frac{f_0+f_1}{2}\right) + h\left(\frac{f_1+f_2}{2}\right) + \cdots + h\left(\frac{f_{n-1}+f_n}{2}\right) \\ &= h\left(\frac{f_0+f_n}{2} + \sum_{i=1}^{n-1} f_i\right).\end{aligned}$$

Obtemos assim uma regra de quadratura da forma (8.1) com pesos

$$\alpha_0 = \alpha_n = \frac{h}{2}, \quad \alpha_i = h,\ i = 1, 2, \ldots n-1$$

dada por

$$Q^{TC}(f,h) = h\left(\frac{f_0+f_n}{2} + \sum_{i=1}^{n-1} f_i\right)$$

a que chama regra dos trapézios composta e cujo valor depende da função integranda f e do espaçamento h. Como veremos no seguinte resultado, a fórmula do erro é semelhante.

Teorema 8.18 (Regra dos trapézios composta).

Seja $f : [a,b] \to \mathbb{R}$ uma função duas vezes diferenciável. Seja ainda

$$Q^{TC}(f,h) = h\left(\frac{f(x_0)+f(x_n)}{2} + \sum_{i=1}^{n-1} f(x_i)\right) \tag{8.11}$$

a regra dos trapézios composta nos nós igualmente espaçados

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + n\,h = b,$$

com espaçamento $h = \frac{b-a}{n}$.

Então o erro da regra dos trapézios composta é dado por

$$I(f) - Q^{TC}(f,h) = -\frac{h^2(b-a)}{12} f''(\xi),$$

para algum $\xi \in [a,b]$. Desta forma temos o majorante para o erro absoluto

$$e^{TC}(f) = \left|I(f) - Q^{TC}(f,h)\right| \leq \frac{h^2(b-a)}{12} \max_{x\in[a,b]} |f''(x)|\,. \tag{8.12}$$

Demonstração. Para demonstrar a fórmula do erro da regra dos trapézios composta, basta considerar a fórmula do erro da regra dos trapézios simples (ver teorema 8.8) em cada um dos sub-intervalos $[x_i, x_{i+1}]$, obtendo

$$I(f) - Q^{TC}(f,h) = -\sum_{i=0}^{n-1} \frac{h^3}{12} f''(\xi_i)$$

para alguns $\xi \in [x_i, x_{i+1}]$. O resultado sai agora do uso do teorema do valor médio (aplicado à segunda derivada de f) e da relação $h = (b-a)/n$. □

Nas regras de quadratura compostas, convém estar atento a ordem de convergência em função do espaçamento h, isto é, tentar perceber quão rapidamente decresce o erro, quando decresce o espaçamento h. temos então a seguinte definição.

Definição 8.19 (Ordem de convergência de uma regra de quadratura).

Diz-se que a regra de quadratura $Q(f,h)$ tem ordem de convergência p se

$$I(f) - Q(f,h) = O(h^p), \quad h \to 0,$$

ou seja, se existe uma constante $C > 0$ e um espaçamento h_0 tal que para $h < h_0$ se tem

$$|I(f) - Q(f,h)| \leq Ch^p.$$

Determinar a ordem de convergência determina a rapidez com que decresce o erro. Na realidade, se uma regra de quadratura tiver ordem p ao decrescer o espaçamento para metade, é esperado que o erro decresça por um fator de $1/2^p$. Intuitivamente, isto pode ser visto pelo facto de, se Q_h tem ordem p então

$$I(f) - Q(f,h) \approx Ch^p, \quad h \approx 0$$

logo obtemos a razão entre os erros

$$\frac{I(f) - Q(f,h/2)}{I(f) - Q(f,h)} \approx \frac{C(h/2)^p}{C(h)^p} \approx \frac{1}{2^p}.$$

Para funções duas vezes diferenciáveis, a regra dos trapézios composta garante um erro menor, quanto menor for o espaçamento h. Nestas condições, podemos escrever a seguinte relação

$$I(f) = Q^{TC}(f,h) + O(h^2),$$

isto é, o erro decresce com h^2 quando $h \to 0$.

Nota 8.20. Devido ao erro ser da ordem de $O(h^2)$, a regra dos trapézios composta tem ordem de convergência 2 para funções duas vezes diferenciáveis em $[a,b]$. Na prática, isto quer dizer que se se diminuir o espaçamento para metade, o erro absoluto decresce com um fator de cerca de $1/4$.

Exercício 8.21. Determine três aproximações para $\int_0^1 e^x\,dx$ pela regra dos trapézios composta, considerando $h = 1, h = 0.5$ e $h = 0.25$. Determine o erro absoluto efetivo da aproximação sabendo que

$$\int_0^1 e^x\,dx \approx 1.71828\ldots$$

Resolução.

O espaçamento $h = 1$ corresponde à regra dos trapézios simples, pelo que temos

$$Q^{TC}(f, 1) = Q^T(f) = h\,\frac{f(0) + f(1)}{2} \approx 1.85914$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f, 1)| = 0.140859.$$

Para $h = 0.5$ temos os 3 nós de quadratura

$$x_0 = 0, \quad x_1 = 0.5, \quad x_2 = 1,$$

logo obtemos a aproximação

$$Q^{TC}(f, 0.5) = h\,\left(\frac{f(0) + f(1)}{2} + f(0.5)\right) \approx 1.75393$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f, 0.5)| = 0.0356493.$$

Por fim, considerando $h = 0.25$ temos os 5 nós de quadratura

$$x_0 = 0, \quad x_1 = 0.25, \quad x_2 = 0.5, \quad x_3 = 0.75, \quad x_4 = 1,$$

logo obtemos a aproximação

$$Q^{TC}(f, 0.25) = h\,\left(\frac{f(0) + f(1)}{2} + f(0.25) + f(0.5) + f(0.75)\right) \approx 1.72722$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f, 0.25)| = 0.00894008.$$

Note-se que quando o espaçamento passou a metade, o erro absoluto diminui com um fator de cerca de $1/4$.

Exercício 8.22. Determine três aproximações para $\int_0^1 \sqrt{x}\,dx$ pela regra dos trapézios composta, considerando $h = 1, h = 0.5$ e $h = 0.25$. Determine o erro absoluto efetivo da aproximação sabendo que

$$\int_0^1 \sqrt{x}\,dx = \frac{2}{3}.$$

Como explica que o decaimento do erro não seja por um fator de $1/4$?

Resolução.

O espaçamento $h = 1$ corresponde à regra dos trapézios simples, pelo que temos

$$Q^{TC}(f,1) = Q^{T}(f) = h\,\frac{f(0)+f(1)}{2} \approx 0.5$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f,1)| = 0.166667.$$

Para $h = 0.5$ temos os 3 nós de quadratura

$$x_0 = 0, \quad x_1 = 0.5, \quad x_2 = 1,$$

logo obtemos a aproximação

$$Q^{TC}(f,0.5) = h\left(\frac{f(0)+f(1)}{2} + f(0.5)\right) \approx 0.603553$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f,0.5)| = 0.0631133.$$

Por fim, considerando $h = 0.25$ temos os 5 nós de quadratura

$$x_0 = 0, \quad x_1 = 0.25, \quad x_2 = 0.5, \quad x_3 = 0.75, \quad x_4 = 1,$$

logo obtemos a aproximação

$$Q^{TC}(f,0.25) = h\left(\frac{f(0)+f(1)}{2} + f(0.25) + f(0.5) + f(0.75)\right) \approx 0.643283$$

o que corresponde a um erro absoluto efetivo

$$|I(f) - Q^{TC}(f,0.25)| = 0.0233836.$$

Note-se que a função $f(x) = \sqrt{x}$ não é diferenciável em $x = 0$, logo a fórmula do erro não é válida, pelo que não se pode esperar que o erro seja da ordem de $O(h^2)$.

Exercício 8.23. Aproxime o valor dos seguintes integrais pela regra dos trapézios composta, usando 1,2 e 4 sub-intervalos de igual comprimento e estabeleça um majorante para o erro absoluto em cada caso.

(a) $\int_0^{\frac{\pi}{4}} \sin(x)\,dx$;

(b) $\int_0^{\frac{\pi}{3}} \cos(x)\,dx$;

(c) $\int_0^1 \sin(x^2)\,dx$;

(d) $\int_0^1 \ln(\cos(x))\,dx$;

Resposta.

(a) Como

$$\max_{x\in[0,\pi/4]} |f''(x)| = \max_{x\in[0,\pi/4]} |-\sin(x)| = \frac{\sqrt{2}}{2},$$

temos

- $h = \frac{\pi}{4} \Rightarrow Q^{TC}(f) = 0.27768,\ e^{TC} \leq 0.028548$;
- $h = \frac{\pi}{8} \Rightarrow Q^{TC}(f) = 0.28912,\ e^{TC} \leq 0.0071370$;
- $h = \frac{\pi}{16} \Rightarrow Q^{TC}(f) = 0.291952,\ e^{TC} \leq 0.0017842$;

(b) Como

$$\max_{x\in[0,\pi/3]} |f''(x)| = \max_{x\in[0,\pi/3]} |-\cos(x)| = 1,$$

temos

- $h = \frac{\pi}{3} \Rightarrow Q^{TC}(f) = 0.785398,\ e^{TC} \leq 0.095698$;
- $h = \frac{\pi}{6} \Rightarrow Q^{TC}(f) = 0.846149,\ e^{TC} \leq 0.023925$;
- $h = \frac{\pi}{12} \Rightarrow Q^{TC}(f) = 0.861073,\ e^{TC} \leq 0.0059811$;

(c) Como

$$\max_{x\in[0,1]} |f''(x)| = \max_{x\in[0,1]} |2\cos(x^2) - 4x^2\sin(x^2)| \leq 6,$$

temos

- $h = 1 \Rightarrow Q^{TC}(f) = 0.420735,\ e^{TC} \leq 0.5$;
- $h = 0.5 \Rightarrow Q^{TC}(f) = 0.33407,\ e^{TC} \leq 0.125$;
- $h = 0.25 \Rightarrow Q^{TC}(f) = 0.315975,\ e^{TC} \leq 0.03125$;

(d) Como

$$\max_{x\in[0,1]} |f''(x)| = \max_{x\in[0,1]} \left|-\frac{1}{\cos^2 x}\right| = \frac{1}{\cos^2(1)} \approx 3.42552,$$

temos

- $h = 1 \Rightarrow Q^{TC}(f) = -0.307813,\ e^{TC} \leq 0.28546$;

- $h = 0.5 \Rightarrow Q^{TC}(f) = -0.219199$, $e^{TC} \leq 0.071365$;
- $h = 0.25 \Rightarrow Q^{TC}(f) = -0.195595$, $e^{TC} \leq 0.0178412$;

Exercício Octave 8.24.

Em Octave, a função `trapz(lx,lf)` determina a aproximação pela regra dos trapézios composta da função f que assume os valores na lista `lf` nos nós da lista `lx`. Utilize esta função em Octave para determinar as aproximações obtidas para os integrais nas secções 8.1.1 e 8.2.1.

8.2.2 Regra de Simpson composta

Da mesma forma, podemos aplicar a regra de Simpson em cada subintervalo, obtendo a regra de Simpson composta. De notar que agora em cada subintervalo precisamos de três nós, para poder aplicar a regra de Simpson. Consideramos então os $n + 1$ nós de quadratura

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + n\,h = b,$$

com espaçamento $h = \frac{b-a}{n}$ e tal que n é par (isto é, temos um número par de sub-intervalos). Da mesma forma, aplicar a regra de Simpson em cada subintervalo da forma $[x_i, x_{i+2}]$, com $i = 0, 2, \ldots, n-2$ é equivalente a considerar o polinómio interpolador de grau 2 em cada um desses sub-intervalos.

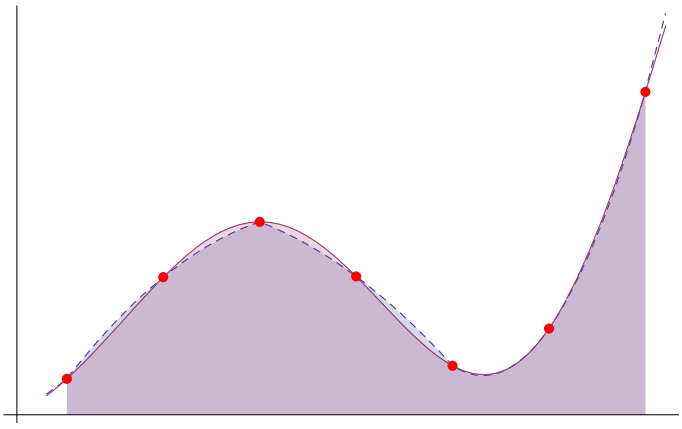


Figura 8.4: *Área sob os gráficos de f (linha contínua) e do polinómio interpolador de grau 2 em cada subintervalo $[x_i, x_i + 2]$ (a tracejado).*

Assim obtemos

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \cdots + \int_{x_{n-2}}^{x_n} f(x)\,dx \\
&\approx h\left(\frac{f_0+4f_1+f_2}{3}\right) + h\left(\frac{f_2+4f_3+f_4}{3}\right) + \cdots + h\left(\frac{f_{n-2}+4f_{n-1}+f_n}{3}\right) \\
&= h\left(\frac{f_0}{3} + \frac{4f_1}{3} + \frac{2f_2}{3} + \frac{4f_3}{3} + \frac{2f_4}{3} + \cdots + \frac{4f_{n-1}}{3} + \frac{f_n}{3}\right) \\
&= \frac{h}{3}\left(f_0 + f_n + 4\sum_{i=0}^{\frac{n}{2}-1} f_{2i+1} + 2\sum_{i=1}^{\frac{n}{2}-1} f_{2i}\right).
\end{aligned}$$

Esta é a regra de Simpson composta, que é uma regra de quadratura da forma (8.1) com pesos

$$\alpha_i = \begin{cases} \frac{h}{3}, & i = 0 \vee i = n, \\ \frac{4h}{3}, & i \text{ ímpar}, \\ \frac{2h}{3}, & i \text{ par} \wedge i \neq 0 \wedge i \neq n \end{cases} \tag{8.13}$$

com $i = 0, 1, \ldots, n$. Comparativamente à regra de Simpson simples, a fórmula do erro é também semelhante neste caso.

Teorema 8.25 (Regra de Simpson composta).
Seja $f : [a, b] \to \mathbb{R}$ uma função quatro vezes diferenciável. Seja ainda

$$Q^{SC}(f, h) = \sum_{i=0}^{n} \alpha_j\, f(x_j) \tag{8.14}$$

a regra de Simpson composta nos $n + 1$ nós (n par) igualmente espaçados

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + n\,h = b,$$

com espaçamento $h = \frac{b-a}{n}$ e pesos de quadratura dados por (8.13).
Então o erro da regra de Simpson composta é dado por

$$I(f) - Q^{SC}(f, h) = -\frac{h^4(b-a)}{180} f^{(4)}(\xi),$$

para algum $\xi \in [a, b]$. Desta forma temos o majorante para o erro absoluto

$$e^{SC}(f) = \left|I(f) - Q^{SC}(f, h)\right| \leq \frac{h^4(b-a)}{180} \max_{x\in[a,b]} \left|f^{(4)}(x)\right|. \tag{8.15}$$

Demonstração. Vem da aplicação da fórmula do erro da regra de Simpson simples em cada um dos sub-intervalos. □

Para funções quatro vezes diferenciáveis, a regra de Simpson composta garante um erro menor, quanto menor for o espaçamento h, tal como no caso da regra dos trapézios composta. No entanto, agora temos

$$I(f) = Q^{SC}(f, h) + O(h^4),$$

isto é, o erro decresce com h^4 quando $h \to 0$, o que nos dá um decrescimento mais rápido que no caso da regra de trapézios composta.

Nota 8.26. Como o erro é da ordem de $O(h^4)$, a regra de Simpson composta tem ordem de convergência 4 para funções quatro vezes diferenciáveis em $[a, b]$. Na prática, isto quer dizer que se se diminuir o espaçamento para metade, o erro absoluto decresce com um fator de cerca de $1/16$.

Exercício 8.27. Aproxime o valor dos seguintes integrais pela regra de Simpson composta, usando 3 e 5 nós de quadratura igualmente espaçados e estabeleça um majorante para o erro absoluto em cada caso.

(a) $\int_0^{\frac{\pi}{4}} \sin(x)\, dx$;

(b) $\int_0^{\frac{\pi}{3}} \cos(x)\, dx$;

(c) $\int_0^1 \sin(x^2)\, dx$;

(d) $\int_0^1 \ln(\cos(x))\, dx$;

Resposta.

(a) Como

$$\max_{x\in[0,\pi/4]} |f^{(4)}(x)| = \max_{x\in[0,\pi/4]} |\sin(x)| = \frac{\sqrt{2}}{2},$$

temos

- $h = \frac{\pi}{8} \Rightarrow Q^{SC}\left(f, \frac{\pi}{8}\right) = 0.292933$, $e^{SC} \leq 0.000073374$;
- $h = \frac{\pi}{16} \Rightarrow Q^{SC}\left(f, \frac{\pi}{16}\right) = 0.292896$, $e^{SC} \leq 4.5859 \times 10^{-6}$;

(b) Como

$$\max_{x\in[0,\pi/3]} |f^{(4)}(x)| = \max_{x\in[0,\pi/3]} |\cos(x)| = 1,$$

temos

- $h = \frac{\pi}{6} \Rightarrow Q^{SC}\left(f, \frac{\pi}{6}\right) = 0.866399$, $e^{SC} \leq 0.00043727$;
- $h = \frac{\pi}{12} \Rightarrow Q^{SC}\left(f, \frac{\pi}{12}\right) = 0.866048$, $e^{SC} \leq 0.000027329$;

(c) Como

$$\max_{x\in[0,1]} |f^{(4)}(x)| = \max_{x\in[0,1]} |-48x^2\cos(x^2) - 12\sin(x^2) + 16x^4\sin(x^2)| \leq 76,$$

temos

- $h = 0.5 \Rightarrow Q^{SC}(f, 0.5) = 0.305181$, $e^{SC} \leq 0.0263889$;
- $h = 0.25 \Rightarrow Q^{SC}(f, 0.25) = 0.309944$, $e^{SC} \leq 0.00164931$;

(d) Como

$$\max_{x\in[0,1]} |f^{(4)}(x)| = \max_{x\in[0,1]} \left|\frac{2 + 4\sin^2(x))}{\cos^2(x)}\right| = \frac{2 + 4\sin^2(1)}{\cos^2(1)} \approx 56.703,$$

temos

- $h = 0.5 \Rightarrow Q^{SC}(f, 0.5) = -0.189661$, $e^{SC} \leq 0.0196885$;
- $h = 0.25 \Rightarrow Q^{SC}(f, 0.25) = -0.187727$, $e^{SC} \leq 0.00123053$;

Exercício Octave 8.28.

Escreva uma função em Octave

`SimpsonComposta(lx,lf)`

que dados os nós `lx` igualmente espaçados no intervalo [a,b] e os respetivos valores `lf` da função f, determina a aproximação do integral $\int_a^b f(x)dx$ pela regra de Simpson composta. Utilize esta função para confirmar as aproximações obtidas no exercício anterior.

Resolução.

No ficheiro `SimpsonComposta.m` escrevemos:

```
function Q = SimpsonComposta(lx,lf)
if max(size(lx))/2 == round(max(size(lx))/2)
  disp('n nao e par!  Nao se pode aplicar Simpson!');
else
  h= lx(2)-lx(1);
  alfa = ones(size(lx));
  alfa(2:2:end-1)=4;
```

```
  alfa(3:2:end-2)=2;
  Q = h*dot(alfa,lf)/3;
end
return
```

Exercício 8.29. Para calcular o valor do integral $\int_0^1 e^{-x^2} dx$ consideraram-se os valores tabelados:

x_i	0	1	2	3	4
$e^{-x_i^2}$	1	0.367879	0.0183156	1.2341×10^{-4}	1.12535×10^{-7}

.

Determine aproximações para o integral usando as regras dos trapézios e Simpson, nas suas versões simples e compostas. Em cada caso determine um majorante para o erro sabendo que

$$\max_{x\in[0,4]} |f^{(n)}(x)| \leq 2(n-1)!, \quad 1 \leq n \leq 5.$$

Resposta.

Temos

$$\begin{aligned} Q^T(f) &= 2.00000, & E^T &\leq 10.667; \\ Q^S(f) &= 0.715508, & E^S &\leq 4.2667; \\ Q^{TC}(f,1) &= 0.886319, & E^{TC} &\leq 0.66667; \\ Q^{SC}(f,1) &= 0.836214, & E^{SC} &\leq 0.26667; \end{aligned}$$

8.3 Quadraturas com nós não-igualmente espaçados

Muitas vezes, em problemas da vida real, temos acesso ao valor de f em vários pontos $x_0, x_1, \ldots x_n$, em que o espaçamento entre os nós não é igual. Isto deve-se ao facto de que nem sempre se tem acesso aos dados quando e onde pretendemos, sendo utilizar os dados disponíveis para estabelecermos as aproximações. Desta forma é importante estabelecer regras de quadratura da forma (8.1)

$$Q_n(f) = \sum_{j=0}^{n} \alpha_j \, f(x_j)$$

em que os nós de quadratura x_j, $j = 0, 1, \ldots, n$ não são igualmente espaçados. Como temos $n+1$ graus de liberdade (correspondentes aos valores dos pesos α_j), geralmente impomos que a regra de quadratura seja exata para polinómios de grau n,

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, a_j \in \mathbb{R},$$

ou, dito de outra forma, obtemos a regra de quadratura por integração exata do polinómio interpolador de grau n nos nós dados. Como os monómios x^j, $j \in \mathbb{N}$ são linearmente independentes, a imposição

$$I(p_n) = Q_n(p_n)$$

reduz-se ao sistema linear

$$\begin{cases} I(1) = Q_n(1) = \alpha_0 + \alpha_1 + \alpha_2 + \cdots + \alpha_n \\ I(x) = Q_n(x) = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \\ I(x^2) = Q_n(x^2) = \alpha_0 x_0^2 + \alpha_1 x_1^2 + \alpha_2 x_2^2 + \cdots + \alpha_n x_n^2 \\ \vdots \\ I(x^n) = Q_n(x^n) = \alpha_0 x_0^n + \alpha_1 x_1^n + \alpha_2 x_2^n + \cdots + \alpha_n x_n^n \end{cases}$$

que podemos escrever na forma matricial na forma

$$\begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & x_2^n & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} I(1) \\ I(x) \\ I(x^2) \\ \vdots \\ I(x^n) \end{bmatrix}.$$

Assim basta resolver o sistema linear anterior, por forma a encontrar os pesos de quadratura α_j, $j = 0, 1, \ldots, n$, e depois aplicar a regra de quadratura nos pontos dados.

Nota 8.30. A uma matriz da forma

$$\begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & x_2^n & \ldots & x_n^n \end{bmatrix}$$

chama-se matriz de *Vandermonde*. Uma matriz de *Vandermonde* é sempre invertível para nós x_j, $j = 0, 1, \ldots, n$ distintos entre si. No entanto o condicionamento da matriz pode deteriorar-se com o aumento da dimensão da matriz, pelo que muitas vezes é necessário recorrer a métodos de regularização.

Exercício 8.31. Conhece-se o valor de uma função f segunda a tabela:

x_i	0	2	6
f_i	1	2	3

.

(a) Determine uma regra de quadratura exata para polinómios de grau 2 nos nós dados para aproximar o integral $I(f) = \int_0^8 f(x)dx$.

(b) Determine uma aproximação para o integral $I(f) = \int_0^8 f(x)dx$, pela regra de quadratura anterior.

Resolução.

(a) Temos os nós de quadratura

$$x_0 = 0, \quad x_1 = 2, \quad x_2 = 6.$$

Assim, queremos determinar os pesos α_j, $j = 0, 1, 2$, de forma a obter uma regra de quadratura

$$Q_n(f) = \sum_{j=0}^{n} \alpha_j f(x_j)$$

exata para polinómios de grau 2. Temos o sistema linear seguinte

$$\begin{bmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} I(1) \\ I(x) \\ I(x^2) \end{bmatrix}.$$

em que

$$\begin{aligned} I(1) &= \int_0^8 1dx = [x]_0^8 = 8; \\ I(x) &= \int_0^8 xdx = \left[\frac{x^2}{2}\right]_0^8 = 32; \\ I(x^2) &= \int_0^8 x^2dx = \left[\frac{x^3}{3}\right]_0^8 = \frac{512}{3} \approx 170.667. \end{aligned}$$

Assim, temos

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 6 \\ 0 & 4 & 36 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 32 \\ \frac{512}{3} \end{bmatrix}.$$

cuja solução é

$$\alpha_0 = \frac{8}{9} \approx 0.888889, \quad \alpha_1 = \frac{8}{3} \approx 2.66667, \quad \alpha_2 = \frac{40}{9} \approx 4.44444.$$

Definimos a regra de quadratura exata para polinómios 2 no intervalo $[0, 8]$ por

$$Q_2(f) = \sum_{j=0}^{2} \alpha_j f(x_j) = \frac{8}{9} f(0) + \frac{8}{3} f(2) + \frac{40}{9} f(6).$$

(b) Temos a aproximação

$$\int_0^8 f(x)\,dx \approx Q_2(f) = \frac{8}{9} \times 1 + \frac{8}{3} \times 2 + \frac{40}{9} \times 3 = \frac{176}{9} \approx 19.5556.$$

Exercício 8.32. Determine uma regra de quadratura exata para polinómios de grau apropriado para aproximar o integral dado, considerando os pontos tabelados.

(a)

x_i	1	2	4
f_i	2	2	3

para o integral $\int_1^4 f(x)\,dx$.

(b)

x_i	1	2	4
f_i	1	0	1

para o integral $\int_1^4 f(x)\,dx$.

(c)

x_i	1	2	4
f_i	1	0	1

para o integral $\int_0^4 f(x)\,dx$.

(d)

x_i	0	2	4
f_i	1	2	1

para o integral $\int_0^4 f(x)\,dx$.

(e)

x_i	0	1	3
f_i	-1	0	2

para o integral $\int_1^3 f(x)\,dx$.

(f)

x_i	0	1	3	4
f_i	-1	0	2	1

para o integral $\int_0^4 f(x)\,dx$.

(g)

x_i	-2	0	1	2
f_i	0	0	1	1

para o integral $\int_{-2}^{2} f(x)\,dx$.

Resposta.

(a) $Q_2(f) = \frac{9}{4}f(2) + \frac{3}{4}f(4), \quad \int_1^4 f(x)\,dx \approx 6.75;$

(b) $Q_2(f) = \frac{9}{4}f(2) + \frac{3}{4}f(4), \quad \int_1^4 f(x)\,dx \approx 0.75;$

(c) $Q_2(f) = \frac{16}{9}f(1) + \frac{4}{3}f(2) + \frac{8}{9}f(4), \quad \int_0^4 f(x)\,dx \approx 2.66667;$

(d) $Q_2(f) = Q^S(f) = \frac{2}{3}f(0) + \frac{8}{3}f(2) + \frac{2}{3}f(4), \quad \int_0^4 f(x)\,dx \approx 6.66667;$

(e) $Q_2(f) = -\frac{4}{9}f(0) + \frac{5}{3}f(1) + \frac{7}{9}f(3), \quad \int_1^3 f(x)\,dx \approx 2;$

(f) $Q_3(f) = \frac{2}{9}f(0) + \frac{16}{9}f(1) + \frac{16}{9}f(3) + \frac{2}{9}f(4), \quad \int_0^4 f(x)\,dx \approx 3.55556;$

(g) $Q_3(f) = \frac{2}{3}f(-2) + \frac{8}{3}f(1) + \frac{2}{3}f(2), \quad \int_0^4 f(x)\,dx \approx 0.666667;$

Exercício Octave 8.33.

Escreva uma função em Octave

`PesosQuadratura(lx,a,b)`

que dados os $n+1$ nós de quadratura `lx` determine os pesos de quadratura de forma a obter uma regra de quadratura exata de grau n para o integral $\int_a^b f(x)dx$. Utilize essa função para verificar os pesos de quadratura obtidos no exercício 8.32.

Resolução.

No ficheiro `PesosQuadratura.m` escrevemos:

```
function alfa = PesosQuadratura(lx,a,b)
n = max(size(lx));
matA = ones(n);
vb = ones(n,1);
vb(1) = b-a;
for ii=1:n-1
  matA(ii+1,:)  = lx.^(ii);
  vb(ii+1) = b^(ii+1)/(ii+1)-a^(ii+1)/(ii+1);
end
alfa = matA\vb;
return
```

Como exemplo, tomando a primeira alínea do exercício 8.32, obtemos o resultado executando o comando `PesosQuadratura([1,2,4],1,4)`.

Exercício Octave 8.34.

Escreva uma função em Octave

`QuadraturaExata(lx,lf,a,b)`

que dados os $n+1$ nós de quadratura `lx` e os respetivos valores `lf` da função f determine a aproximação para o integral $\int_a^b f(x)dx$ por uma regra de quadratura exata de grau n. Utilize essa função para verificar os resultados do exercício 8.32.

Resolução.

Utilizando a função do exercício anterior, no ficheiro `QuadraturaExata.m` escrevemos:

```
function Q = QuadraturaExata(lx,lf,a,b)
alfa = PesosQuadratura(lx,a,b);
Q = dot(alfa,lf);
return
```

Como exemplo, tomando a primeira alínea do exercício 8.32, obtemos o resultado executando o comando `QuadraturaExata([1,2,4],[2,2,3],1,4)`.

Capítulo 9

Métodos Numéricos para equações diferenciais ordinárias

Nesta última secção vamo-nos concentrar em métodos numéricos para resolver equações diferencias ordinárias. Começamos então por definir formalmente e classificar equações diferenciais.

Uma equação diferencial é uma equação que envolve derivadas da função que é sua solução. Desta forma, existem várias classes de equações diferenciais e várias formas de classificação. Uma equação diferencial pode ser classificada quanto ao **tipo**, da seguinte forma:

- **Equação diferencial ordinária (EDO)** se envolver apenas derivadas totais da solução (ou seja, soluções que dependam de apenas uma variável);
- **Equação diferencial às derivadas parciais (EDP)** se envolver derivadas parciais da solução (ou seja, soluções que dependam de duas ou mais variáveis).

Note-se que a solução de uma EDP será uma função de duas ou mais variáveis, enquanto que a solução de uma EDO será à partida uma função de apenas uma variável.

Uma equação diferencial pode também ser classificada quanto à **ordem** e quanto ao **grau**. Assim, uma equação diferencial diz-se de **ordem n** se a maior ordem de derivada envolvida for n e diz-se de **grau p** se a potência da derivada de maior ordem for p.

Neste capítulo vamos apenas considerar equações diferenciais ordinárias, deixando o tratamento numérico para obtenção de soluções de EDP para o capítulo 10. Assim formalizamos na seguinte definição mais alguns conceitos relativos a este tipo de equações.

Definição 9.1 (Equação Diferencial Ordinária (EDO)).

Seja $y = y(t)$ uma função $n-$vezes diferenciável na variável real t.

Seja ainda $F = F\left(t, y, y', y'', \ldots, y^{(n)}\right)$ uma função real dependente de t, de y e das derivadas de y em t de ordem inferior ou igual a $n \in \mathbb{N}$.

Chama-se **equação diferencial ordinária (EDO) de ordem n** a uma equação do tipo

$$F\left(t, y, y', y'', \ldots, y^{(n)}\right) = 0$$

onde a solução $y = y(t)$ é uma função de t. A EDO diz-se **escalar** se a sua solução y for uma função real e diz-se **vetorial** se a sua solução y for uma função vetorial. Caso a solução da EDO dependa de constantes arbitrárias e represente todas as soluções da EDO, diz-se uma **solução geral**.

Chama-se **condição inicial** a qualquer condição adicional da forma $y^{(j)}(t_0) = y_j$ para algum $j \in \{0, 1, \ldots, n-1\}$. Ao problema

$$\begin{cases} F(t, y, y', y'', \ldots, y^{(n)}) = 0 \\ y(t_0) = y_0 \\ y'(t_0) = y_1 \\ \quad \vdots \\ y^{(n-1)}(t_0) = y_{n-1} \end{cases}$$

chama-se **problema de valores inicial (PVI)** associado à EDO

$$F(t, y, y', y'', \ldots, y^{(n)}) = 0.$$

A designação de valor inicial vem do facto de geralmente se ter $t_0 = 0$.

Por simplicidade, começamos por considerar problemas de valor inicial (PVI) associados a EDO de primeira ordem, ou seja, da forma

$$\begin{cases} y' = f(t, y), \quad t \geq t_0 \\ y(t_0) = y_0. \end{cases} \tag{9.1}$$

As EDOs de primeira ordem envolvem apenas derivadas de primeira ordem da solução. Desta forma, elas aparecem geralmente para modelar taxas de variação da solução em função da sua variável de dependência.

A primeira questão que nos devemos colocar quando em presença de uma equação é se ela tem solução (existência de solução) e se a solução, caso exista, é única (unicidade de solução). O teorema seguinte dá-nos algumas condições para as quais é possível estabelecer existência e unicidade de solução de EDOs de primeira ordem.

Teorema 9.2 (Picard-Lindelöf).

Sejam $f = f(t, y)$ e a sua derivada parcial $\frac{\partial f}{\partial y}$ em y duas funções contínuas num retângulo $\mathcal{R}$ (fechado) contendo (t_0, y_0). Então o problema de valor inicial (PVI)

$$\begin{cases} y' = f(t, y) \\ y(t_0) = y_0 \end{cases}$$

tem uma única solução em $\mathcal{R}$.

Exercício 9.3. Indique se pode garantir existência e unicidade de solução para $t \in \mathbb{R}$ para os PVIs seguintes

(a) $\begin{cases} y' = \sin(yt) \\ y(0) = 1 \end{cases}$

(b) $\begin{cases} y^2 y' = e^t \\ y(0) = 0 \end{cases}$

(c) $\begin{cases} y' = \dfrac{yt}{y^2 + 1} \\ y(1) = 0 \end{cases}$

Resposta.

(a) Sim.

(b) Não.

(c) Sim.

Note-se que caso a EDO envolvida seja linear e de coeficientes constantes, já mencionamos no capítulo 1 que a resolução pode ser feita analiticamente [1, 3]. Caso contrário, não sabemos em geral encontrar uma solução analítica, pelo que temos de optar por um método numérico que nos aproxime a solução. Muitas das vezes o processo analítico de resolução é muito complexo, pelo que a opção por um método numérico de aproximação é nesses casos também a melhor escolha.

O método de Euler (que introduziremos de seguida) baseia-se na expansão em fórmula de Taylor de $y = y(t)$. Dado um instante $t = T$, o objetivo é aproximar o valor de $y(T)$.

9.1 Método de Euler

O Método de Euler baseia-se na expansão em polinómio de Taylor de grau 1 da solução da EDO e pode ser aplicado em vários contextos, nomeadamente, para aproximar soluções de EDO escalares de primeira ordem e de EDO vetoriais de primeira ordem, ou seja, em que a solução é um função vetorial de uma única variável. Este último permitirá obter aproximações de soluções de EDO escalares de ordem superior.

Assim dividimos a abordagem nestes três casos, começando pelo mais simples, que depois se generalizará para os restantes.

9.1.1 Método de Euler para EDOs escalares de primeira ordem

Seja $y = y(t)$ uma função real de variável real, que é solução do PVI

$$\begin{cases} y'(t) = f\Big(t, y(t)\Big), \quad t \geq t_0 \\ y(t_0) = y_0, \end{cases}$$

e pretende-se aproximar o valor de $y(T)$, para algum instante T dado. Supondo que y é duas vezes diferenciável em t, temos pela formula de Taylor

$$y(t+h) = y(t) + \underbrace{y'(t)}_{=f(t,y(t))} h + O(h^2).$$

Assim, para h pequeno, temos

$$y(t+h) \approx y(t) + h\, f(t, y(t)). \tag{9.2}$$

Consideramos agora os $n+1$ nós igualmente espaçados no intervalo $[t_0, T]$

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_0 + 2h,\, \ldots,\, t_n = t_0 + n\, h = T$$

com espaçamento $h = \frac{T - t_0}{n}$ suficientemente pequeno. Assim, para obter uma aproximação de $y(t_1)$ podemos utilizar o valor

$$y_1 := y(t_0) + h\, f(t_0, y(t_0)) = y_0 + h\, f(t_0, y_0)$$

e de (9.2) temos $y(t_1) \approx y_1$. De seguida, podemos aproximar o valor de $y(t_2)$ utilizando a aproximação de $y(t_1)$, pois de (9.2) temos

$$y(t_2) = y(t_1 + h) \approx y(t_1) + h\, f(t_1, y(t_1)) \approx y_1 + h\, f(t_1, y_1) =: y_2$$

e temos $y(t_2) \approx y_2$. Definimos então o procedimento iterativo

$$y_{j+1} = y_j + h\, f(t_j, y_j), \quad j = 0, 1, \ldots n-1$$

e obtemos a aproximação

$$y(T) = y(t_n) \approx y_n.$$

A este método chama-se método de Euler.

Definição 9.4 (Método de Euler para EDOs escalares de 1ª ordem).

Seja $y : [t_0, T] \to \mathbb{R}$ a solução duas vezes diferenciável, i.e. $y \in C^2([t_0, T])$ do PVI

$$\begin{cases} y' = f(t, y), \quad t \in [t_0, T], \\ y(t_0) = y_0. \end{cases}$$

Então o valor de $y(T)$, para um dado $T \geq t_0$ pode ser aproximado por

$$y(T) \approx y_n$$

em que y_n é obtido pelo **método de Euler**

$$\begin{cases} y_0 \equiv \text{valor da condição inicial} \\ y_{j+1} = y_j + h\, f(t_j, y_j). \end{cases} \tag{9.3}$$

para $j = 0, 1, \ldots n-1$, em que os $n+1$ nós

$$t_0,\ t_1 = t_0 + h,\ t_2 = t_0 + 2h,\ \ldots,\ t_n = t_0 + n\,h = T$$

são igualmente espaçados, com espaçamento $h = \frac{T - t_0}{n}$ suficientemente pequeno.

Tal como no caso das regras de quadratura para integração numérica, convém estabelecer a ordem de convergência do método, isto é, estudar como o erro decresce com o decrescimento do espaçamento h considerado. Temos o resultado seguinte.

Teorema 9.5 (Ordem de Convergência do método de Euler).

Seja $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ uma função Lipschitz na segunda variável, isto é, existe uma constante $L > 0$ tal que

$$|f(t, x) - f(t, y)| \leq L|x - y|. \tag{9.4}$$

Então o método de Euler (9.3) *converge para a solução y duas vezes diferenciável do PVI* (9.1) *com convergência linear, isto é, para $h = (T - t_0)/n$ e*

$$t_0,\ t_1 = t_0 + h,\ t_2 = t_1 + h, \ldots,\ t_n = t_0 + nh = T$$

temos

$$e_{n,h} := |y(T) - y_n| = O(h), \quad h \to 0.$$

Demonstração. Pela fórmula de Taylor, temos para y duas vezes diferenciável que

$$\begin{aligned} y(T) = y(t_n) &= y(t_{n-1}) + y'(t_{n-1})h + y''(\xi)\frac{h^2}{2} \\ &= y(t_{n-1}) + hf(t_{n-1}, y(t_{n-1})) + y''(\xi)\frac{h^2}{2}. \end{aligned}$$

Pela definição do método de Euler temos

$$y_n = y_{n-1} + hf(t_{n-1}, y_{n-1}).$$

Assim obtemos

$$e_{n,h} := |y(T) - y_n| \leq |y(t_{n-1}) - y_{n-1}| + h|f(t_{n-1}, y(t_{n-1})) - f(t_{n-1}, y_{n-1})| + |y''(\xi)|\frac{h^2}{2}.$$

Como f é Lipschitz na segunda variável, temos

$$|f(t_{n-1}, y(t_{n-1})) - f(t_{n-1}, y_{n-1})| \leq L|y(t_{n-1})), y_{n-1}|.$$

Por outro lado, como y é duas vezes diferenciável, pelo Teorema de Weierstrass existe $M > 0$ tal que

$$\max_{t \in t_0, T} |y''(t)| = M.$$

Assim concluímos que

$$\begin{aligned} e_{n,h} &\leq (1 + Lh)|y(t_{n-1}) - y_{n-1}| + M\frac{h^2}{2} \\ &\leq (1 + Lh)e_{n-1,h} + M\frac{h^2}{2}. \end{aligned}$$

Repetindo o processo $n - 1$ vezes, temos

$$\begin{aligned} e_{n,h} &\leq (1 + Lh)|y(t_{n-1}) - y_{n-1}| + \frac{Mh^2}{2} \\ &\leq (1 + Lh)^{n-1}e_{0,h} + (1 + (1 + Lh) + (1 + Lh)^2 + \cdots + (1 + Lh)^n)\frac{Mh^2}{2}. \end{aligned}$$

Pela soma de uma série geométrica

$$\sum_{i=0}^{n} A^n = \frac{A^n - 1}{A - 1}$$

obtemos

$$e_{n,h} \leq (1+Lh)^n e_{0,h} + \frac{(1+Lh)^n - 1}{(1+Lh) - 1}\frac{Mh^2}{2}$$
$$\leq (1+Lh)^n e_{0,h} + \frac{(1+Lh)^n - 1}{L}\frac{Mh}{2}.$$

Notando agora que a série de Taylor da função exponencial é dada por

$$e^x = \sum_{i=0}^{\infty} \frac{x^n}{n!},$$

temos

$$e^{Lh} = \sum_{n=0}^{\infty} \frac{(Lh)^n}{n!} \geq \sum_{n=0}^{1} \frac{(Lh)^n}{n!} = 1 + Lh, \tag{9.5}$$

logo

$$e_{n,h} \leq e^{Lhn} e_{0,h} + \frac{e^{Lhn} - 1}{L}\frac{Mh^2}{2}$$
$$\leq e^{L(T-t_0)} e_{0,h} + \frac{e^{L(T-t_0)} - 1}{L}\frac{Mh^2}{2} \tag{9.6}$$

uma vez que $T - t_0 = nh$. Como pela condição inicial temos $y_0 = y(t_0)$, temos $e_{0,h} = 0$, o que termina a demonstração. □

Nota 9.6. Um vez que o método de Euler tem apenas ordem de convergência 1, para um espaçamento pequeno, espera-se que se o espaçamento diminuir para metade, o erro diminua também para cerca de metade. Desta forma, a ordem de convergência obtida numericamente por

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

deve ser próxima de 1 e tender para este valor à medida que o espaçamento h diminui.

Nota 9.7. A estimativa de erro (9.6) ilustra também a instabilidade e mau condicionamento do método de Euler. De facto, se o erro inicial não for nulo, este é amplificado por um fator de $e^{L(T-t_0)}$, o que pode ser catastrófico para T muito maior que t_0 ($T \gg t_0$) ou para L grande. Isto mostra em particular que os erros de arredondamento podem ser amplificados com o decorrer das iterações.

Exercício 9.8. Aproxime a solução do PVI

$$\begin{cases} y' = y + e^t \cos(t), \quad t \geq 0 \\ y(0) = 0 \end{cases}$$

no ponto $t = 1$, utilizando os espaçamentos $h = 1$, $h = 0.5$ e $h = 0.25$. Sabendo que a solução exata é dada por $y(t) = e^t \sin t$, determine os erros absolutos cometidos em cada caso e comente o seu decrescimento.

Resposta.

Neste caso, temos

$$x_0 = 0,\ y_0 = 0,\ T = 1,\ f(t, y) = y + e^t \cos(t).$$

Assim, considerando o espaçamento $h = 1$, temos a aproximação

$$y(1) \approx y_1 := y_0 + h\, f(t_0, y_0) = e^0 \cos 0 = 1.$$

O erro absoluto cometido é $|y(1) - y_1| = |e^1 \sin(1) - 1| = 1.28736$.

Considerando o espaçamento $h = 0.5$, temos a aproximação

$$\begin{cases} y(0.5) \approx y_1 := y_0 + h\, f(t_0, y_0) = 0.5(e^0 \cos 0) = 0.5 \\ y(1) \approx y_2 := y_1 + h\, f(t_1, y_1) = 0.5 + 0.5(0.5 + e^{0.5} \cos(0.5)) = 1.47344 \end{cases}$$

O erro absoluto cometido é $|y(1) - y_2| = |e \sin(1) - 1.47344| = 0.813911$.

Considerando finalmente o espaçamento $h = 0.25$, temos a aproximação

$$\begin{cases} y(0.25) \approx y_1 := y_0 + h\, f(t_0, y_0) = 0.25(e^0 \cos 0) = 0.25 \\ y(0.5) \approx y_2 := y_1 + h\, f(t_1, y_1) = 0.25 + 0.25(0.25 + e^{0.25} \cos(0.25)) = 0.623527 \\ y(0.75) \approx y_3 := y_2 + h\, f(t_2, y_2) = 0.6235 + 0.25(0.6235 + e^{0.5} \cos(0.5)) = 1.14113 \\ y(1) \approx y_4 := y_3 + h\, f(t_3, y_3) = 1.14113 + 0.25(1.14113 + e^{0.75} \cos(0.75)) = 1.81366 \end{cases}$$

O erro absoluto cometido é $|y(1) - y_4| = |e \sin(1) - 1.47344| = 0.473696$. Neste último passo, quando o espaçamento h diminuiu para metade, o erro parece diminuir para cerca de metade. Esta relação seria mais notória se continuássemos o procedimento para espaçamentos mais pequenos.

Exercício 9.9. Aproxime a solução dos PVIs seguintes no ponto $t = 2$, utilizando n sub-intervalos, com $n = 1, n = 2$ e $n = 4$.

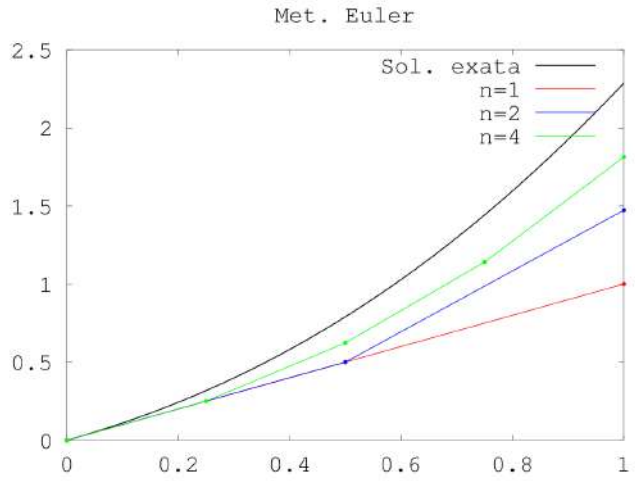


Figura 9.1: *As três aproximações do exercício 9.8, graficamente.*

(a) $\begin{cases} y' = ty, \quad t \geq 1 \\ y(1) = 2 \end{cases}$

(b) $\begin{cases} y' = y^t, \quad t \geq 0 \\ y(0) = 0.1 \end{cases}$

(c) $\begin{cases} y' = t\sin((t-1)y), \quad t \geq 0 \\ y(0) = -1 \end{cases}$

Resposta.

(a) Temos

$$\begin{cases} n = 1 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_1 = 4; \\ n = 2 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_2 = 5.25; \\ n = 4 \Rightarrow h = 0.25 & \Rightarrow \quad y(2) \approx y_4 = 6.4856. \end{cases}$$

(b) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = 2.1; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = 2.2; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = 2.3821; \end{cases}$$

(c) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = -1; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = -1; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = -1.1996; \end{cases}$$

Exercício Octave 9.10. Determine uma função Octave

```
MetEuler(expf,t0,y0,T, n)
```

que dadas a expressão `expf` de f em função de t e y, o instante inicial `t0`, a condição inicial `y0`, o instante final `T` e o número de subdivisões `n` do intervalo a considerar, determine as aproximações pelo método de Euler da solução do problema de valor inicial 9.1 nos instantes

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T,$$

com espaçamento $h = (T - t_0)/\mathrm{n}$. Utilize-o para verificar os resultados obtidos nesta secção.

Resposta.

No ficheiro `MetEuler.m` escrevemos os comandos:

```
function ly = MetEuler(expf,t0,y0,T, n)
h=(T-t0)/n;
f=inline(expf,'t','y');
lt=t0:h:T;
ly=zeros(size(lt));
ly(1)=y0;
for j=1:n
  ly(j+1)=ly(j)+h*f(lt(j),ly(j));
end
return
```

9.1.2 Método de Euler para EDOs vetoriais de primeira ordem

Da mesma forma, o método de Euler também se pode aplicar a EDOs de 1ª ordem vetoriais. Seja então $Y = Y(t)$ uma função vetorial com m componentes da forma

$$Y(t) = \begin{bmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_m(t) \end{bmatrix}$$

em que a notação Y_j representa agora a j-ésima componente do vetor Y. Seja $Y = Y(t)$ solução do PVI

$$\begin{cases} Y' = F(t, Y), \quad t \geq t_0 \\ Y(t_0) = Y^{[0]}. \end{cases}$$

Note-se que as igualdades são entre vetores pois a função F é agora uma função vetorial, dada por (com um ligeiro abuso de notação em F)

$$F(t, Y_1, Y_2, \ldots, Y_m) = \begin{bmatrix} F_1(t, Y_1, Y_2, \ldots, Y_m) \\ F_2(t, Y_1, Y_2, \ldots, Y_m) \\ \vdots \\ F_m(t, Y_1, Y_2, \ldots, Y_m) \end{bmatrix}$$

Assim,

$$\begin{cases} Y_1' = F_1(t, Y_1, Y_2, \ldots, Y_m) \\ Y_2' = F_2(t, Y_1, Y_2, \ldots, Y_m) \\ \vdots \\ Y_m' = F_m(t, Y_1, Y_2, \ldots, Y_m) \end{cases}, \quad t \geq t_0,$$

enquanto que $Y(t_0) = Y^{[0]}$ se traduz na igualdade componente a componente

$$\begin{cases} Y_1(t_0) = Y_1^{[0]} \\ Y_2(t_0) = Y_2^{[0]} \\ \vdots \\ Y_m(t_0) = Y_m^{[0]} \end{cases}$$

em que o vetor correspondente à iterada inicial $Y^{[0]}$ é dado por

$$Y^{[0]} = \begin{bmatrix} Y_1^{[0]} \\ Y_2^{[0]} \\ \vdots \\ Y_m^{[0]} \end{bmatrix}.$$

Podemos então aplicar o método de Euler, da seguinte forma.

Definição 9.11 (Método de Euler para EDO vetoriais de 1ª ordem).
Seja $Y : [t_0, T] \to \mathbb{R}^n$ a solução vetorial duas vezes diferenciável em t do PVI

$$\begin{cases} Y' = F(t, Y), \quad t \geq t_0 \\ Y(t_0) = Y^{[0]}. \end{cases}$$

Então o valor de $Y(T)$, para um dado $T \geq t_0$ pode ser aproximado por

$$Y(T) \approx Y^{[n]}$$

em que $Y^{[n]}$ é obtido pelo método de Euler vetorial

$$\begin{cases} Y^{[0]} \equiv \text{valor da condição inicial} \\ Y^{[j+1]} = Y^{[j]} + h\, F(t_j, Y^{[j]}). \end{cases} \tag{9.7}$$

para $j = 0, 1, \ldots n-1$, em que os $n+1$ nós

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_0 + 2h,\, \ldots,\, t_n = t_0 + n\,h = T$$

são igualmente espaçados, com espaçamento $h = \frac{T-t_0}{n}$ suficientemente pequeno. De forma similar, $Y^{[j]}$ pode ser visto como uma aproximação de $Y(t_j)$, isto é,

$$Y(t_j) \approx Y^{[j]}.$$

Nota 9.12. Mostra-se de forma semelhante ao caso escalar que o método de Euler vetorial tem convergência linear para soluções $Y = Y(t)$ duas vezes diferenciável e $F = F(t, Y)$ Lipschitz na segunda derivada.

Assim sendo, a aplicação do método de Euler para funções vetoriais é semelhante ao caso escalar, notando apenas que a solução é agora uma função vetorial.

Exercício 9.13. Determine uma aproximação do valor de $x(1)$, em que x satisfaz o PVI

$$\begin{cases} x'(t) = e^t + (y(t))^2 \\ y'(t) = (z(t))^2 - t\sin(x(t)) \\ z'(t) = tx(t)y(t) \end{cases}$$

que verifica $x(0) = 0$, $y(0) = 0$, e $z(0) = 1$, utilizando o método de Euler com espaçamento $h = 0.5$.

Resolução.

Neste caso, temos $t_0 = 0$, $T = 1$ e

$$Y(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}, \quad Y^{[0]} = \begin{bmatrix} x(0) \\ y(0) \\ z(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad F(t,x,y,z) = \begin{bmatrix} e^t + y^2 \\ z^2 - t\,\sin(x) \\ t\,x\,y \end{bmatrix}.$$

Assim, usando espaçamento $h = 0.5$, obtemos

$$Y(0.5) = \begin{bmatrix} x(0.5) \\ y(0.5) \\ z(0.5) \end{bmatrix} \approx Y^{[1]} = Y^{[0]} + h\underbrace{F(t_0, Y_1^{[0]}, Y_2^{[0]}, Y_3^{[0]}}_{=F(0,0,0,1)} = \begin{bmatrix} 0.5 \times e^0 \\ 0.5 \times 1^2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \end{bmatrix}.$$

Repetindo o processo, temos

$$Y(1) = \begin{bmatrix} x(1) \\ y(1) \\ z(1) \end{bmatrix} \approx Y^{[2]} = Y^{[1]} + h \underbrace{F(t_1, Y_1^{[1]}, Y_2^{[1]}, Y_3^{[1]})}_{=F(0.5,0.5,0.5,1)} = \begin{bmatrix} 1.44936 \\ 0.88014 \\ 1.0625 \end{bmatrix}.$$

Assim obtemos a aproximação $x(1) \approx 1.44936$.

Exercício 9.14. Determine uma aproximação do valor de $z(2)$, em que z satisfaz o PVI

$$\begin{cases} x'(t) = (z(t) + y(t))^2 \\ y'(t) = t(x(t) - z(t))^2 \\ z'(t) = (tx(t) + y(t))^2 \end{cases}$$

que verifica $x(0) = 1$, $y(0) = -1$, e $z(0) = 1$, utilizando o método de Euler com espaçamento $h = 1$.

Resolução.

Neste caso, temos $t_0 = 0$, $T = 2$ e

$$Y(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}, \quad Y^{[0]} = \begin{bmatrix} x(0) \\ y(0) \\ z(0) \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \quad F(t, x, y, z) = \begin{bmatrix} (z + y)^2 \\ t(x - z)^2 \\ (tx + y)^2 \end{bmatrix}.$$

Assim, usando espaçamento $h = 1$, obtemos

$$Y(1) = \begin{bmatrix} x(1) \\ y(1) \\ z(1) \end{bmatrix} \approx Y^{[1]} = Y^{[0]} + h \underbrace{F(t_0, Y_1^{[0]}, Y_2^{[0]}, Y_3^{[0]})}_{=F(0,1,-1,1)} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}.$$

Repetindo o processo, temos

$$Y(2) = \begin{bmatrix} x(2) \\ y(2) \\ z(2) \end{bmatrix} \approx Y^{[2]} = Y^{[1]} + h \underbrace{F(t_1, Y_1^{[1]}, Y_2^{[1]}, Y_3^{[1]})}_{=F(1,1,-1,2)} = \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}.$$

Assim obtemos a aproximação $z(2) \approx 2$.

9.1.3 Método de Euler para EDOs escalares de ordem n.

Suponhamos agora que temos um PVI com uma EDO de ordem n da forma

$$\begin{cases} y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \\ \qquad y(t_0) = y_0, \\ \qquad y'(t_0) = y_1, \\ \qquad \vdots \\ \qquad y^{(n-1)}(t_0) = y_{n-1}. \end{cases} \tag{9.8}$$

para uma função escalar $f = f(t, y, y', \ldots, y^{(n-1)})$ que depende da variável temporal t e da função $y = y(t)$ e de todas as suas derivada em t até à ordem $n-1$. É possível também obter uma aproximação da solução num instante T através do método de Euler. Para isso é necessário reduzir a ordem da EDO escalar para uma EDO vetorial de primeira ordem. Definimos assim o vetor

$$Y := \begin{bmatrix} y \\ y' \\ y'' \\ \vdots \\ y^{(n-1)} \end{bmatrix}$$

ou seja temos as componentes do vetor Y dadas por

$$Y_1 = y, \quad Y_2 = y', \quad Y_3 = y'', \quad \ldots Y_n = y^{(n-1)}. \tag{9.9}$$

Tomando a derivada de Y e o facto da $j-$ésima componente de Y ser a derivada de ordem $j-1$ de y, temos

$$Y' = \begin{bmatrix} y' \\ y'' \\ \vdots \\ y^{(n-1)} \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} Y_2 \\ Y_3 \\ \vdots \\ Y_n \\ y^{(n)} \end{bmatrix}$$

Tomando a EDO de (9.8) dada por

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$$

obtemos a EDO de primeira ordem

$$Y' = F(t, Y) = F\left(t, [Y_1, Y_2, \ldots, Y_n]^T\right)$$

em que a função vetorial F é dada por

$$F(t,Y) := \begin{bmatrix} Y_2(t) \\ Y_3(t) \\ \vdots \\ Y_n(t) \\ f(t,Y_1,Y_2,\ldots,Y_n) \end{bmatrix}$$

uma vez que por definição de Y, temos a relação (9.9).

Além disso, da condição inicial de (9.8), temos a condição de inicial para o vetor Y dada por

$$Y^{[0]} := Y(t_0) = \begin{bmatrix} y(t_0) \\ y'(t_0) \\ y''(t_0) \\ \vdots \\ y^{(n-1)}(t_0) \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{bmatrix}.$$

Assim o PVI (9.8) é equivalente ao PVI vetorial

$$\begin{cases} Y' = F(t,Y), \\ Y(t_0) = Y^{[0]}. \end{cases} \tag{9.10}$$

Assim podemos aplicar o método de Euler para funções vetoriais (9.7) ao PVI anterior para aproximar a solução y, uma vez que a função y é a primeira componente da função vetorial Y.

Exercício 9.15. Determine uma aproximação de $y(1)$ pelo método de Euler com espaçamento $h = 0.5$, sabendo que y é solução do PVI

$$\begin{cases} y'' + 2y' + y = \cos 2t, \\ y(0) = 0, \\ y'(0) = 0. \end{cases}$$

Compare a aproximação com o valor da solução exata

$$y(t) = \frac{3}{25}\, e^{-t} - \frac{1}{5}\, t\, e^{-t} - \frac{3}{25}\cos(2t) + \frac{4}{25}\sin(2t).$$

Resposta.

Definimos a função vetorial Y por

$$Y(t) := \begin{bmatrix} y(t) \\ y'(t) \end{bmatrix}.$$

Assim temos

$$Y' = \begin{bmatrix} y' \\ y'' \end{bmatrix}$$

e como

$$y'' = \cos(2t) - 2y' - y = \cos(2t) - 2Y_2 - Y_1$$

temos EDO vetorial

$$Y' = F(t, Y)$$

com

$$F(t, Y) = \begin{bmatrix} Y_2 \\ \cos(2t) - 2Y_2 - Y_1 \end{bmatrix}.$$

Temos também a condição inicial

$$Y^{[0]} = Y(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Assim, aplicando o método de Euler com $h = 0.5$, temos

$$t_0 = 0, \quad t_1 = 0.5, \quad t_2 = 1,$$

e logo obtemos

$$\begin{aligned} Y(0.5) \approx Y^{[1]} &= Y^{[0]} + h\, F(t_0, Y^{[0]}) \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.5 \begin{bmatrix} 0 \\ \cos(2 \times 0) - 2 \times 0 - 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}. \end{aligned}$$

e

$$\begin{aligned} Y(1) \approx Y^{[2]} &= Y^{[1]} + h\, F(t_1, Y^{[1]}) \\ &= \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + 0.5 \begin{bmatrix} 0.5 \\ \cos(2 \times 0.5) - 2 \times 0.5 - 0 \end{bmatrix} \\ &= \begin{bmatrix} 0.25 \\ 0.270151 \end{bmatrix}. \end{aligned}$$

Assim, temos a aproximação $y(1) \approx Y_1^{[2]} = 0.25$. Comparando com o valor exato $y(1) = 0.165994\ldots$ obtemos os erros absoluto e relativo de

$$e = |y(1) - Y_1^{[2]}| = 0.084005 \quad \text{e} \quad \delta = \frac{|y(1) - Y_1^{[2]}|}{|y(1)|} = 0.506070$$

ou seja, um erro de cerca de 50%. Para melhorar a aproximação haveria que reduzir o passo h do método de Euler.

Exercício 9.16. Determine uma aproximação de $y(3)$ pelo método de Euler com espaçamento $h = 1$, sabendo que y é solução do PVI

$$\begin{cases} y''' = y\,y' + e^{-t\,y''}, \\ y(1) = -1, \\ y'(1) = 1 \\ y''(1) = 0. \end{cases}$$

Resposta.

Definimos a função vetorial Y por

$$Y(t) := \begin{bmatrix} y(t) \\ y'(t) \\ y''(t) \end{bmatrix}.$$

Assim temos

$$Y' = F(t, Y)$$

para

$$F(t, Y) = \begin{bmatrix} Y_2 \\ Y_3 \\ Y_1 \times Y_2 + e^{-t\,Y_3} \end{bmatrix}.$$

Temos também a condição inicial

$$Y^{[0]} = Y(1) = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}.$$

Aplicando o método de Euler com $h = 1$, temos

$$t_0 = 1, \quad t_1 = 2, \quad t_2 = 3,$$

e logo obtemos

$$Y(2) \approx Y^{[1]} = Y^{[0]} + h\,F(t_0, Y^{[0]}) = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + 1 \times \begin{bmatrix} 1 \\ 0 \\ -1 \times 1 + e^{-0\times 0} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

e

$$Y(3) \approx Y^{[2]} = Y^{[1]} + h\,F(t_1, Y^{[1]}) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1 \times \begin{bmatrix} 1 \\ 0 \\ 1 \times 0 + e^{-0\times 0} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Assim, temos a aproximação $y(3) \approx Y_1^{[2]} = 1$.

9.2 Métodos de Runge-Kutta

Os métodos de Runge-Kutta (RK) pretendem melhorar a ordem de convergência do método de Euler. Vamos estudá-los no contexto de EDOs escalares de primeira ordem, se bem que possam ser generalizados para EDos vetoriais de 1ª ordem ou EDOs escalares de ordem superior, à semelhança do que foi feito para o método de Euler nas secções 9.1.2 e 9.1.3, respetivamente.

Os métodos RK partem também da expansão em série de Taylor da solução $y = y(t)$ do problema de valor inicial (9.1), mas agora pelo menos de terceira ordem (em vez de segunda como no método de Euler), ou seja,

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + O(h^3), \quad h \to 0. \tag{9.11}$$

Assim, precisamos da segunda derivada de $y = y(t)$, sendo que do problema de valor inicial (9.1) temos $y' = f(t, y(t))$, ou seja, pela regra da derivada composta obtemos

$$\begin{aligned} y''(t) &= \frac{d}{dt} f(t, y(t)) \\ &= \frac{\partial f}{\partial x}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\, y'(t) \\ &= \frac{\partial f}{\partial x}(t, y(t)) + f(t, y(t)) \frac{\partial f}{\partial y}(t, y(t)) \end{aligned} \tag{9.12}$$

em que $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial y}$ são as derivadas parciais de f em relação à primeira e segunda variável.

Nota 9.17. Como é evidente, assumimos que a solução y é três vezes diferenciável para a expansão de Taylor, pelo que as derivadas parciais de f existem.

A ideia base dos métodos de RK passa agora por encontrar outro tipo de representações que não necessitem do cálculo explícito das derivadas parciais de f e não comprometam a ordem de convergência.

Pela fórmula de Taylor para funções de duas variáveis, obtemos que

$$f(t+h_t, y+h_y) = f(t,y) + h_t \frac{\partial f}{\partial x}(t, y(t)) + h_y \frac{\partial f}{\partial x}(t, y(t)) + O(h_t h_y)$$

logo obtemos

$$Ahf(t,y) = A\,hf(t+h_t, y+h_y) - A\,h\,h_t \frac{\partial f}{\partial x}(t, y(t)) - A, h\,h_y \frac{\partial f}{\partial x}(t, y(t)) + O(h\,h_t h_y). \tag{9.13}$$

para uma contante arbitrária $A \in \mathbb{R}$. Assim, notando que $y = y(t)$ para facilitar a notação, de (9.1), (9.12) e (9.13) concluímos que

$$\begin{aligned}
&hy' + \frac{h^2}{2} y'' = \\
&= hf(t,y) + \frac{h^2}{2}\frac{\partial f}{\partial x}(t,y) + \frac{h^2}{2} f(t,y)\frac{\partial f}{\partial y}(t,y) \\
&= (1-A)\,h\,f(t,y) + A\,h\,f(t,y) + \frac{h^2}{2}\frac{\partial f}{\partial x}(t,y) + \frac{h^2}{2} f(t,y)\frac{\partial f}{\partial y}(t,y) \\
&= (1-A)\,h\,f(t,y) + A\,h\,f(t+h_t, y+h_y) + h\left(\frac{h}{2} - A\,h_t\right)\frac{\partial f}{\partial x}(t,y) \\
&\qquad + h\left(\frac{h}{2} f(t,y) - A\,h_y\right)\frac{\partial f}{\partial y}(t,y) + O(h\,h_t h_y).
\end{aligned} \tag{9.14}$$

Procuramos valores de A, h_t e h_y de forma a garantir que os termos das derivadas parciais se anulem e que o resto é de ordem superior a 2, ou seja, pelo menos $O(h^3)$. Assim, substituindo a expressão (9.14) consoante os vários valores dos parâmetros escolhidos em (9.11), obtemos vários métodos de Runge-Kutta com ordem de convergência (pelo menos) quadrática.

9.2.1 Método RK do ponto médio

O caso mais simples é o chamado método RK do ponto médio, que resulta da escolha

$$A = 1, \quad h_t = \frac{h}{2}, \quad h_y = \frac{h}{2} f(t,y)$$

originando de (9.14) que

$$hy' + \frac{h^2}{2}y'' = hf(t + h/2, y + h\, f(t, y)/2) + O(h^3). \tag{9.15}$$

Pela substituição em (9.11) obtemos

$$y(t + h) = y(t) + hf(t + h/2, y + h\, f(t, y)/2) + O(h^3), \quad h \to 0,$$

que dá o denominado método de Runge-Kutta do ponto médio, que tem convergência quadrática. Este método chama-se do ponto médio, porque em vez de avaliar f em $(t_k, y(t_k))$ (como o método de Euler) para obter a aproximação no instante t_{k+1}, avalia a função f de forma aproximada no instante intermédio $t_k + h/2 = (t_k + t_{k+1})/2$.

Definição 9.18 (Método RK do Ponto médio).

Seja $y : [t_0, T] \to \mathbb{R}$ a solução três vezes diferenciável, i.e. $y \in C^3([t_0, T])$ do PVI

$$\begin{cases} y' = f(t, y), \quad t \in [t_0, T] \\ y(t_0) = y_0. \end{cases}$$

Então o valor de $y(T)$, para um dado $T \geq t_0$ pode ser aproximado por

$$y(T) \approx y_n$$

em que y_n é obtido pelo **método RK do ponto médio**

$$\begin{cases} y_0 \equiv \text{valor da condição inicial} \\ y_{j+1} = y_j + h\, f\left(t_j + \dfrac{h}{2}, y_j + \dfrac{h}{2} f_j\right). \end{cases} \tag{9.16}$$

em que $f_j = f(t_j, y_j)$ para $j = 0, 1, \ldots n - 1$, em que os $n + 1$ nós

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_0 + 2h,\, \ldots,\, t_n = t_0 + n\, h = T$$

são igualmente espaçados, com espaçamento $h = \frac{T - t_0}{n}$ suficientemente pequeno.

Em comparação com o método de Euler, o método RK do ponto-médio tem ordem de convergência superior, pelo que as aproximações serão melhores. No entanto para garantir essa ordem de convergência é necessário que f seja diferenciável, enquanto que no caso de Euler bastava que fosse Lipschitz na segunda variável.

Teorema 9.19 (Ordem de Convergência do método RK do ponto médio).

Seja $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ uma função diferenciável em ambas as variáveis.

Então o método RK do ponto médio (9.16) *converge para a solução y três vezes diferenciável do PVI* (9.1) *com convergência quadrática, isto é, para $h = (T - t_0)/n$ e*

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T$$

temos

$$e_{n,h} := |y(T) - y_n| = O(h^2), \quad h \to 0.$$

Demonstração. Seguimos os passo da demonstração do Teorema 9.5. Temos pelo desenvolvimento em série de Taylor de y

$$y(T) = y(t_n) = y(t_{n-1}) + y'(t_{n-1})h + y''(t_{n-1})\frac{h^2}{2} + y'''(\xi)\frac{h^3}{6}$$

e pela definição do método RK do ponto médio (9.16) que

$$\begin{aligned} y(t_n) - y_n =& y(t_{n-1}) + y'(t_{n-1})h + y''(t_{n-1})\frac{h^2}{2} + y'''(\xi)\frac{h^3}{6} \\ & - y_{n-1} - hf(t_{n-1} + h/2, y_{n-1} + h\, f_{n-1}/2). \end{aligned}$$

Assim, para o erro da aproximação y_n de $y(T)$ dado por $e_{n,h} := |y(t_n) - y_n|$ temos pela desigualdade de Cauchy-Schwartz que

$$\begin{aligned} e_{n,h} \leq & e_{n-1,h} + \left| y'(t_{n-1})h + y''(t_{n-1})\frac{h^2}{2} - hf\left(t_{n-1} + \frac{h}{2}, y(t_{n-1}) + \frac{h}{2}f(t_{n-1}, y(t_{n-1}))\right)\right| \\ & + h\left| f\left(t_{n-1} + \frac{h}{2}, y(t_{n-1}) + \frac{h}{2}f(t_{n-1}, y(t_{n-1}))\right) - f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2}f_{n-1}\right)\right| \\ & + \frac{Mh^3}{6} \end{aligned} \tag{9.17}$$

em que

$$M := \max_{t \in [t_0, T]} |y'''(t)|.$$

Agora, de (9.15) temos

$$\left| y'(t_{n-1})h + y''(t_{n-1})\frac{h^2}{2} - hf\left(t_{n-1} + \frac{h}{2}, y(t_{n-1}) + \frac{h}{2}f(t_{n-1}, y(t_{n-1}))\right)\right| \leq Ch^3$$

e como f é diferenciável na segunda variável e (pelo teorema de Lagrange 6.4) logo Lipschitz com $L = \max_y \left|\frac{\partial f}{\partial y}(t,y)\right|$ que

$$\begin{aligned}
&\left| f\left(t_{n-1} + \frac{h}{2}, y(t_{n-1}) + \frac{h}{2} f(t_{n-1}, y(t_{n-1}))\right) - f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} f_{n-1}\right)\right| \leq \\
&\leq L\left(|y(t_{n-1}) - y_{n-1}| + \frac{h}{2}\left|f(t_{n-1}, y(t_{n-1})) - f_{n-1}\right|\right) \\
&\leq L\left(e_{n-1,h} + \frac{h}{2}\left|f(t_{n-1}, y(t_{n-1})) - f(t_{n-1}, y_{n-1})\right|\right) \\
&\leq L\left(e_{n-1,h} + \frac{h}{2} L\left|y(t_{n-1}) - y_{n-1}\right|\right) \\
&\leq \left(L + \frac{h}{2}L^2\right) e_{n-1,h}
\end{aligned}$$

Assim, utilizando os majorantes anteriores em (9.17) obtemos

$$e_{n,h} \leq \left(1 + Lh + \frac{h^2}{2}L^2\right) e_{n-1,h} + \tilde{C}h^3$$

para uma constante $\tilde{C}$. Repetindo os últimos passos da demonstração do teorema 9.5, obtemos

$$\begin{aligned}
e_{n,h} &\leq \left(1 + Lh + \frac{h^2}{2}L^2\right)^n e_{0,h} + \frac{\left(1 + Lh + \frac{h^2}{2}L^2\right)^n - 1}{\left(1 + Lh + \frac{h^2}{2}L^2\right) - 1}\tilde{C}h^3 \\
&= (1 + Lh)^n e_{0,h} + \frac{\left(1 + Lh + \frac{h^2}{2}L^2\right)^n - 1}{L + \frac{h}{2}L^2}\tilde{C}h^2
\end{aligned}$$

e do majorante $e^{Lh} \geq 1 + Lh + L^2h^2/2$ (ver polinómio de Taylor de segunda ordem da exponencial (9.5)) e do facto de $T - t_0 = nh$, temos

$$e_{n,h} \leq e^{L(T-t_0)} e_{0,h} + \frac{e^{L(T-t_0)} - 1}{L}\tilde{C}h^2.$$

Assim, como $e_{0,h} = |y_(t_0) - y_0| = 0$ pela condição inicial, temos

$$e_{n,h} = O(h^2)$$

terminando a demonstração. □

Nota 9.20. O método RK do ponto médio (9.16) tem ordem de convergência 2. Assim, quando o espaçamento h passa a metade, espera-se que o erro decresça com um fator de $1/4$, partindo do princípio que a solução é suficientemente regular, neste caso, três-vezes diferenciável. Desta forma, a ordem de convergência obtida numericamente por

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

deve ser próxima de 2 e tender para este valor à medida que o espaçamento h diminui.

Exercício 9.21. Determine as aproximações pelo método RK do ponto médio nas condições do exercício 9.8.

Resposta.

Temos as aproximações da tabela seguinte, que se traduzem graficamente nas aproximações da figura 9.2:

t_j	$y(t_i)$			
	n=1	n=2	n=4	Sol. Exata
0	0	0	0	0
0.25	-	-	0.31233	0.31767
0.5	-	0.74705	0.77752	0.79044
0.75	-	-	1.42018	1.44303
1	1.94689	2.16932	2.25243	2.28736

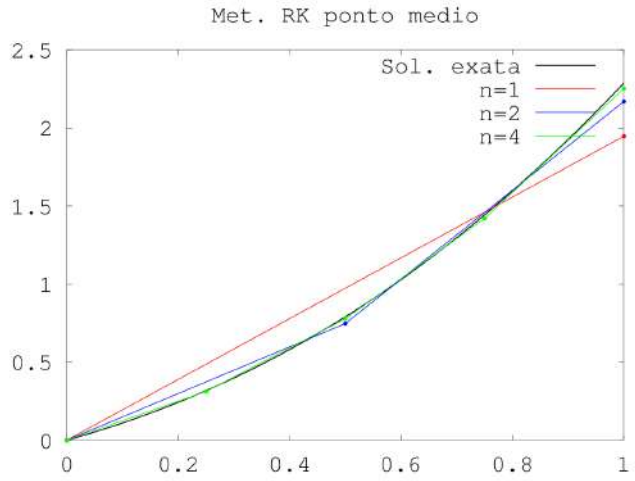


Figura 9.2: *As três aproximações do exercício 9.21, graficamente.*

Temos também os erros e a respetiva ordem de convergência numérica obtida pela expressão

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

na tabela seguinte:

h	$e_h = \lvert y(1) - y_n \rvert$	p
1.00	0.340466	1.528
0.50	0.118039	1.757
0.25	0.034927	-

Como esperado pelo teorema 9.19, a ordem de convergência calculada numericamente tende para $p = 2$, sendo este aspeto mais visível quando o espaçamento h se torna mais próximo de zero.

Exercício 9.22. Determine as aproximações pelo método RK do ponto médio nas condições do exercício 9.9.

Resposta.

(a) Temos

$$\begin{cases} n = 1 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_1 = 6.5; \\ n = 2 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_2 = 7.8486; \\ n = 4 \Rightarrow h = 0.25 & \Rightarrow \quad y(2) \approx y_4 = 8.5701. \end{cases}$$

(b) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = 2.3; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = 2.3772; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = 2.7870; \end{cases}$$

(c) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = -1; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = -1.3169; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = -1.6238; \end{cases}$$

Exercício Octave 9.23. Determine uma função Octave

```
MetRKPontoMedio(expf,t0,y0,T, n)
```

que dadas a expressão `expf` de f em função de t e y, o instante inicial `t0`, a condição inicial `y0`, o instante final `T` e o número de subdivisões `n` do intervalo

a considerar, determine as aproximações pelo método RK do ponto médio da solução do problema de valor inicial 9.1 nos instantes

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T,$$

com espaçamento $h = (T - t_0)/\mathrm{n}$. Utilize-o para verificar os resultados obtidos nesta secção.

Resposta.

No ficheiro `MetRKPontoMedio.m` escrevemos os comandos:

```
function ly = MetRKPontoMedio(expf,t0,y0,T, n)
h=(T-t0)/n;
f=inline(expf,'t','y');
lt=t0:h:T;
ly=zeros(size(lt));
ly(1)=y0;
for j=1:n
  fj=f(lt(j),ly(j));
  ly(j+1)=ly(j)+h*f(lt(j)+h/2,ly(j)+h/2*fj);
end
return
```

9.2.2 Método RK de Euler modificado

Nesta secção vamos ilustrar outra hipótese de escolha de parâmetros de forma a obter um método de Runge-Kutta de ordem 2.

Assim, a partir de (9.14) e com a escolha

$$A = 1/2, \quad h_t = h, \quad h_y = h\, f(t, y)$$

temos

$$hy' + \frac{h^2}{2}y'' = \frac{h}{2}\left[f(t,y) + f\left(t + h, y + hf(t,y)\right)\right] + O(h^3). \tag{9.18}$$

Pela substituição em (9.11) obtemos

$$y(t+h) = y(t) + \frac{h}{2}\left[f(t,y) + f\left(t + h, y + hf(t,y)\right)\right] + O(h^3), \quad h \to 0,$$

que dá o denominado método de Runge-Kutta de Euler modificado. A denominação vem do facto de este método ser aproximadamente a média entre o método de Euler explicito (considerado na secção anterior) e o método de Euler implícito, isto é, em que f é avaliada no instante corrente.

Definição 9.24 (Método RK de Euler modificado).

Seja $y : [t_0, T] \to \mathbb{R}$ a solução três vezes diferenciável, i.e. $y \in C^3([t_0, T])$, do PVI

$$\begin{cases} y' = f(t, y), \quad t \in [t_0, T] \\ y(t_0) = y_0. \end{cases}$$

Então o valor de $y(T)$, para um dado $T \geq t_0$ pode ser aproximado por

$$y(T) \approx y_n$$

em que y_n é obtido pelo **método RK de Euler modificado**

$$\begin{cases} y_0 \equiv \text{valor da condição inicial} \\ y_{j+1} = y_j + \dfrac{h}{2} \left[f_j + f\left(t_{j+1}, y_j + h f_j\right) \right]. \end{cases} \tag{9.19}$$

em que $f_j = f(t_j, y_j)$ para $j = 0, 1, \ldots n-1$, em que os $n+1$ nós

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_0 + 2h,\, \ldots,\, t_n = t_0 + n\,h = T$$

são igualmente espaçados, com espaçamento $h = \frac{T - t_0}{n}$ suficientemente pequeno.

O método RK de Euler modificado tem também ordem de convergência quadrática.

Teorema 9.25 (Ordem de Convergência do método RK de Euler Modificado).

Seja $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ uma função diferenciável em ambas as variáveis.

Então o método RK de Euler modificado (9.19) *converge para a solução y três vezes diferenciável do PVI* (9.1) *com convergência quadrática, isto é, para $h = (T - t_0)/n$ e*

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T$$

temos

$$e_{n,h} := |y(T) - y_n| = O(h^2), \quad h \to 0.$$

Demonstração. Segue os mesmos passos da demonstração do teorema 9.19. □

Nota 9.26. O método RK de Euler modificado (9.19) tem ordem de convergência 2. Assim, quando o espaçamento h passa a metade, espera-se que o erro decresça com um fator de $1/4$, partindo do princípio que a solução é suficientemente regular, neste caso, três-vezes diferenciável. Desta forma, a ordem de convergência obtida numericamente por

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

deve ser próxima de 2 e tender para este valor à medida que o espaçamento h diminui.

Exercício 9.27. Determine as aproximações pelo método RK de Euler modificado nas condições do exercício 9.8.

Resposta.

Temos as aproximações da tabela seguinte, que se traduzem graficamente nas aproximações da figura 9.3:

t_j	$y(t_i)$			
	n=1	n=2	n=4	Sol. Exata
0	0	0	0	0
0.25	-	-	0.31176	0.31767
0.5	-	0.73672	0.77470	0.79044
0.75	-	-	1.41228	1.44303
1	1.73435	2.10693	2.23510	2.28736

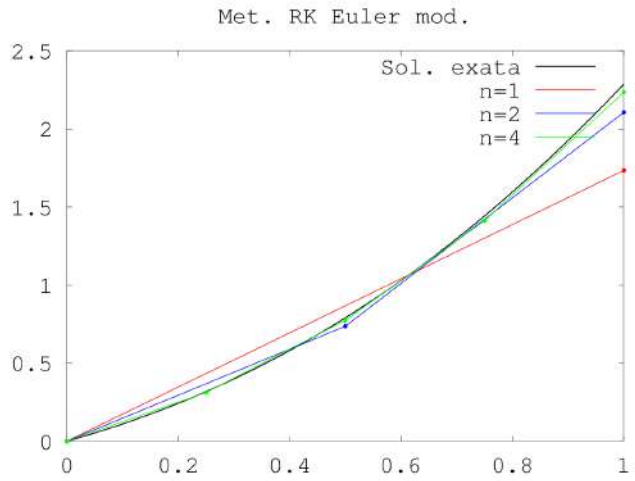


Figura 9.3: *As três aproximações do exercício 9.27, graficamente.*

Temos também os erros e a respetiva ordem de convergência numérica obtida pela expressão

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

na tabela seguinte:

h	$e_h = \lvert y(1) - y_n \rvert$	p
1.00	0.553008	1.616
0.50	0.180425	1.788
0.25	0.052251	-

Como esperado pelo teorema 9.25, a ordem de convergência calculada numericamente tende para $p = 2$, sendo este aspeto mais visível quando o espaçamento h se torna mais próximo de zero.

Exercício 9.28. Determine as aproximações pelo método RK de Euler modificado nas condições do exercício 9.9.

Resposta.

(a) Temos

$$\begin{cases} n = 1 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_1 = 7.0; \\ n = 2 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_2 = 8.1563; \\ n = 4 \Rightarrow h = 0.25 & \Rightarrow \quad y(2) \approx y_4 = 8.6974. \end{cases}$$

(b) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = 5.51; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = 4.37; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = 3.9104; \end{cases}$$

(c) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = -2.6829; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = -1.8415; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = -1.7249; \end{cases}$$

Exercício Octave 9.29. Determine uma função Octave

```
MetRKEulerModificado(expf,t0,y0,T, n)
```

que dadas a expressão `expf` de f em função de t e y, o instante inicial `t0`, a condição inicial `y0`, o instante final `T` e o número de subdivisões `n` do intervalo a considerar, determine as aproximações pelo método RK de Euler modificado da solução do problema de valor inicial 9.1 nos instantes

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T,$$

com espaçamento $h = (T - t_0)/\mathrm{n}$. Utilize-o para verificar os resultados obtidos nesta secção.

Resposta.

No ficheiro `MetRKEulerModificado.m` escrevemos os comandos:

```
function ly = MetRKEulerModificado(expf,t0,y0,T, n)
h=(T-t0)/n;
```

```
f=inline(expf,'t','y');
lt=t0:h:T;
ly=zeros(size(lt));
ly(1)=y0;
for j=1:n
  fj=f(lt(j),ly(j));
  ly(j+1)=ly(j)+h/2*(fj+f(lt(j)+h,ly(j)+h*fj));
end
return
```

9.2.3 Método de RK de ordem 4

Se em vez de se desenvolver o polinómio de Taylor da solução y até à segunda ordem (9.11) como nas secções anteriores, o fizermos para ordens superiores podemos obter métodos RK com ordens de convergência superior.

De notar que nesses casos os cálculos são bem mais morosos que nas secções anteriores, uma vez que implicam derivadas de ordens superiores da função f. Como exemplo, deixamos a expressão de um **método RK de ordem 4** dado por

$$\begin{cases} y_0 \equiv \text{valor da condição inicial} \\ y_{j+1} = y_j + \dfrac{Y_1 + 2Y_2 + 2Y_3 + Y_4}{6} \end{cases} \tag{9.20}$$

em que

$$\begin{aligned} Y_1 &= h\, f(t_j, y_j), \\ Y_2 &= h\, f\left(t_j + \frac{h}{2}, y_j + \frac{Y_1}{2}\right), \\ Y_3 &= h\, f\left(t_j + \frac{h}{2}, y_j + \frac{Y_2}{2}\right), \\ Y_4 &= h\, f\left(t_j + h, y_j + Y_3\right). \end{aligned}$$

Nota 9.30. O método RK (9.20) tem ordem de convergência 4. Assim, quando o espaçamento h passa a metade, espera-se que o erro decresça com um fator de 1/16, partindo do princípio que a solução é suficientemente regular. Desta forma, a ordem de convergência obtida numericamente por

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

deve ser próxima de 4 e tender para este valor à medida que o espaçamento h diminui.

Nota 9.31. Obviamente que qualquer método RK aqui descrito pode ser aplicado a EDO vetoriais de 1ª ordem, por analogia ao que foi feito para o método de Euler na secção 9.1.2. Da mesma forma, analogamente ao que foi feito na secção 9.1.3 para o método de Euler, estes métodos também podem ser generalizados para EDO escalares de ordem superior.

Exercício 9.32. Determine as aproximações pelo método RK de ordem 4 nas condições do exercício 9.8.

Resposta.

Temos as aproximações da tabela seguinte, que se traduzem graficamente nas aproximações da figura 9.4:

t_j	$y(t_i)$			
	n=1	n=2	n=4	Sol. Exata
0	0	0	0	0
0.25	-	-	0.31766	0.31767
0.5	-	0.78993	0.79040	0.79044
0.75	-	-	1.44296	1.44303
1	2.27058	2.28579	2.28724	2.28736

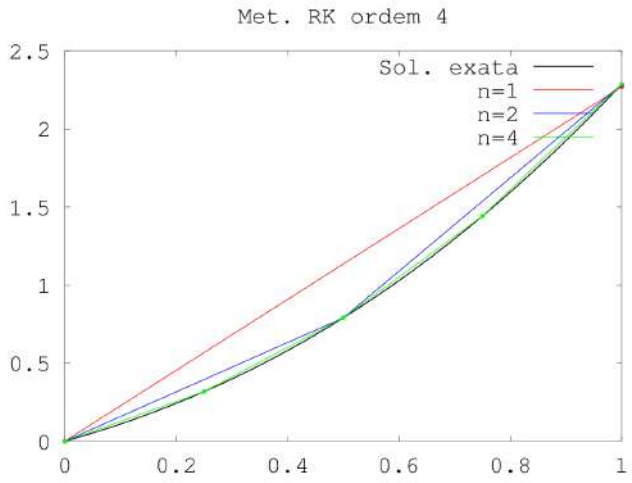


Figura 9.4: *As três aproximações do exercício 9.32, graficamente.*

Temos também os erros e a respetiva ordem de convergência numérica obtida

pela expressão

$$p \approx \log_2 \left(\frac{|e_h|}{|e_{h/2}|} \right)$$

na tabela seguinte:

h	$e_h = \lvert y(1) - y_n \rvert$	p
1.00	0.016777	3.424
0.50	0.001563	3.720
0.25	0.000119	-

Como esperado pela nota 9.31, a ordem de convergência calculada numericamente tende para $p = 4$, sendo este aspeto mais visível quando o espaçamento h se torna mais próximo de zero.

Exercício 9.33. Determine as aproximações pelo método RK de ordem 4 nas condições do exercício 9.9.

Resposta.

(a) Temos

$$\begin{cases} n = 1 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_1 = 8.75; \\ n = 2 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_2 = 8.9374; \\ n = 4 \Rightarrow h = 0.25 & \Rightarrow \quad y(2) \approx y_4 = 8.9611. \end{cases}$$

(b) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = 4.05; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = 3.8027; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = 3.8708; \end{cases}$$

(c) Temos

$$\begin{cases} n = 1 \Rightarrow h = 2 & \Rightarrow \quad y(2) \approx y_1 = -1.5610; \\ n = 2 \Rightarrow h = 1 & \Rightarrow \quad y(2) \approx y_2 = -1.6603; \\ n = 4 \Rightarrow h = 0.5 & \Rightarrow \quad y(2) \approx y_4 = -1.7011; \end{cases}$$

As aproximações pelos quatro métodos considerados da solução para o problema 9.9 podem ser comparadas graficamente na figura 9.5. forma considerados até $n = 16$ subintervalos, para se ter uma melhor noção da convergência do método. Como se pode ver, a aproximação pelo método RK de ordem 4 parece ser a que converge mais rapidamente, seguindo-se os restantes métodos RK (de ordem 2) considerados e finalmente o método de Euler (ordem 1).

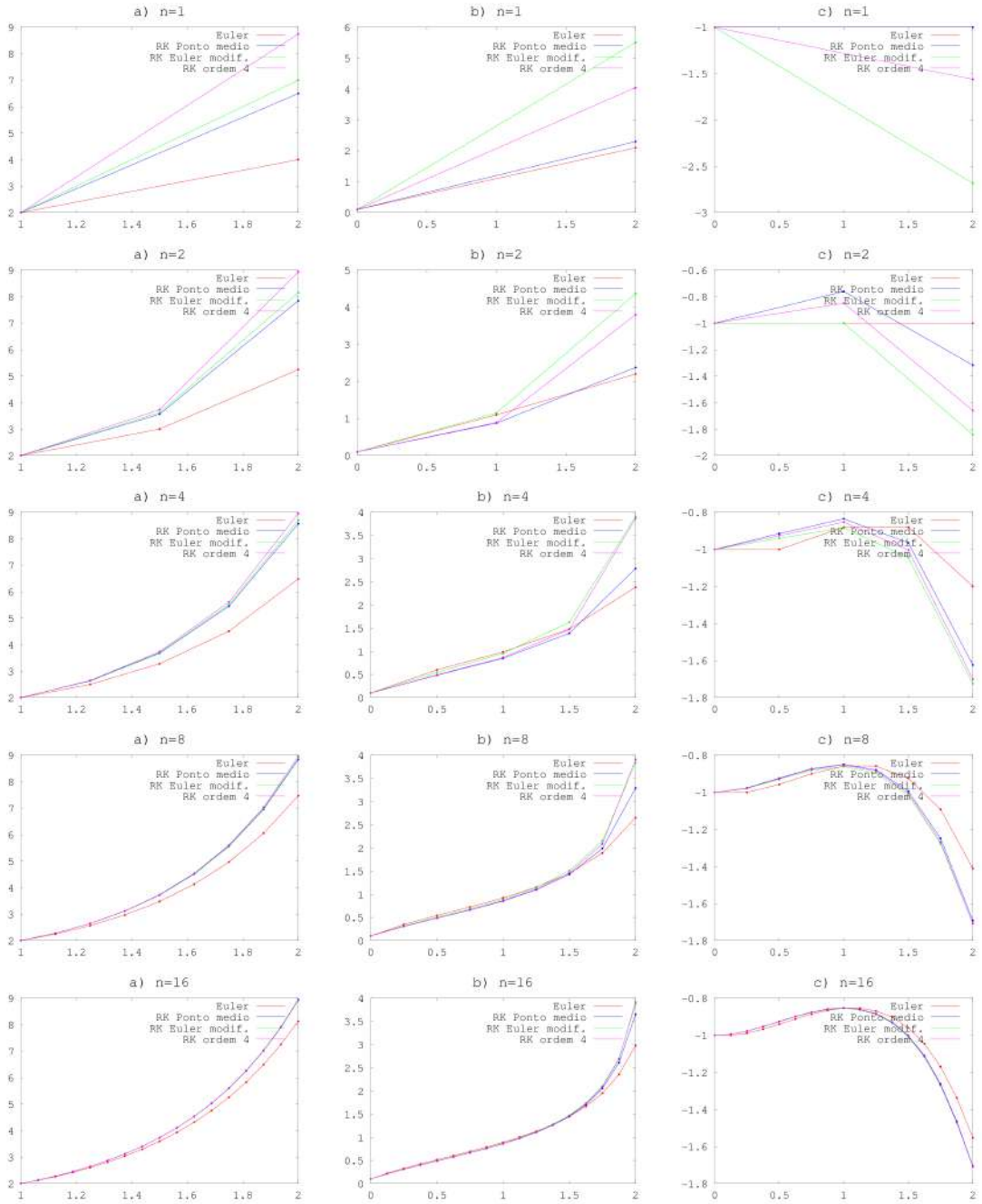


Figura 9.5: *Aproximações da solução do problema 9.9 pelos quatro métodos descritos consoante o número de subintervalos n considerados.*

Exercício Octave 9.34. Determine uma função Octave

```
MetRK4(expf,t0,y0,T, n)
```

que dadas a expressão `expf` de f em função de t e y, o instante inicial `t0`, a condição inicial `y0`, o instante final `T` e o número de subdivisões `n` do intervalo a considerar, determine as aproximações pelo método RK de ordem 4 da solução do problema de valor inicial 9.1 nos instantes

$$t_0,\, t_1 = t_0 + h,\, t_2 = t_1 + h, \ldots,\, t_n = t_0 + nh = T,$$

com espaçamento $h = (T - t_0)/\mathrm{n}$. Utilize-o para verificar os resultados obtidos nesta secção.

Resposta.

No ficheiro `MetRK4.m` escrevemos os comandos:

```
function ly = MetRK4(expf,t0,y0,T, n)
h=(T-t0)/n;
f=inline(expf,'t','y');
lt=t0:h:T;
ly=zeros(size(lt));
ly(1)=y0;
for j=1:n
  fj=f(lt(j),ly(j));
  Y1=h*fj;
  Y2=h*f(lt(j)+h/2,ly(j)+Y1/2);
  Y3=h*f(lt(j)+h/2,ly(j)+Y2/2);
  Y4=h*f(lt(j)+h,ly(j)+Y3);
  ly(j+1)=ly(j)+(Y1+2*Y2+2*Y3+Y4)/6;
end
return
```

Exercício 9.35. Considere o PVI

$$\begin{cases} y' = f(t,y), \\ y(0) = 0 \end{cases}$$

com $f(t,y) = |y-1| + 1/(e-1)$ para $t, y \geq 0$.

(a) Mostre que a função f é Lipschitz na segunda variável para $y \geq 0$.

(b) Justifique que a função f não é diferenciável na segunda variável.

(c) Determine aproximações para $y(1.05)$ pelos quatro métodos estudados, considerando $n = 8$, $n = 16$ e $n = 32$ subintervalos do intervalo $[0, 1.05]$.

(d) Determine a ordem de convergência numérica de cada método neste caso, sabendo que a solução do PVI é

$$y(t) = \begin{cases} \dfrac{e - e^{1-t}}{e-1}, & t \leq 1 \\ \dfrac{e^{t-1} + e - 2}{e-1}, & t > 1 \end{cases}$$

(e) Como justifica os resultados obtidos.

Resposta.

(a) Temos

$$|f(t,x) - f(t,y)| = ||x-1| - |y-1||.$$

Assim, se $x, y \geq 1$, temos

$$|f(t,x) - f(t,y)| = |x - 1 - (y-1)| = |x-y|,$$

se $x, y \leq 1$ (e positivos), temos

$$|f(t,x) - f(t,y)| = |1 - x - (1-y)| = |x-y|.$$

e finalmente se $x \geq 1$ e $y \leq 1$ (e positivo) temos pela desigualdade de Cauchy-Schwartz que

$$\begin{aligned} |f(t,x) - f(t,y)| &= ||x-1| - |y-1|| \\ &\leq |x-1| + |y-1| \\ &= x - 1 + (1-y) \\ &= x - y \\ &= |x-y|. \end{aligned}$$

Assim, a função f é Lipschitz na segunda variável para $y \geq 0$ com constante de Lipschitz $L = 1$.

(b) Como o módulo não é diferenciável na origem, a função f não é diferenciável em $y = 1$.

(c) Utilizando os algoritmos dos exercícios 9.10, 9.23, 9.29 e 9.34, temos a tabela seguinte com as aproximações para cada método e cada espaçamento.

n	Euler	RK P. méd.	RK Eul. mod.	RK ord. 4
8	1.0686916447	1.0265385360	1.0307374076	1.0296289808
16	1.0493536576	1.0292321039	1.0298623490	1.0299008368
32	1.0393692823	1.0295758568	1.0298902352	1.0297938923

(d) Temos também os erros das aproximações e os valores da ordem de convergência numérica na tabela 9.1.

n	Euler		RK P. méd.		RK Eul. mod.		RK ord. 4	
	$\lvert y(T)-y_n\rvert$	p	$\lvert y(T)-y_n\rvert$	p	$\lvert y(T)-y_n\rvert$	p	$\lvert y(T)-y_n\rvert$	p
8	0.0388531	1.0	0.00330005	2.4	0.000898824	5.2	0.000209603	1.8
16	0.0195151	1.0	0.00060648	1.2	2.37651×10^{-5}	-1.1	6.2253×10^{-5}	0.5
32	0.0095307	-	0.000262727	-	5.16513×10^{-5}	-	4.46916×10^{-5}	-

Tabela 9.1: *Erro e ordem de convergência numérica do exercício 9.35.*

(e) Como a função é Lipschitz, está garantida a convergência linear pelo método de Euler pelo teorema 9.5, o que é ilustrado na tabela. No entanto, para os métodos RK considerados a ordem de convergência teórica apenas é válida se a função f for diferenciável (para o método RK de ordem 4, a premissa de suavidade de f é maior) por forma a que a solução y seja duas vezes diferenciável. Note-se inclusivé que no caso do método RK de Euler modificado, o erro aumenta de $n = 16$ para $n = 32$. Assim, a ordem de convergência para os métodos RK é inferior do que a estudada teoricamente com a premissa de diferenciabilidade de f.

Terminamos este capítulo com o uso de resolução numérica de EDOs num modelo aplicado.

Exercício 9.36 (Poluição). Um tanque contém 100 m^3 de água, nos quais estão inicialmente dissolvidos y_0 kg de um agente poluidor. Entra no tanque água contaminada à taxa de 10 m^3 por minuto com uma concentração de poluente de $c(t)$ no instante t, em que

$$\text{Concentração do poluente} = \frac{\text{Massa de poluente}}{\text{Volume}}.$$

Ao entrar no tanque o líquido é uniformemente misturado e sai do tanque à taxa de 10 m^3/min.

(a) Usando a *Lei do Balanço*,

$$\text{Taxa de variação} = \text{Taxa de entrada} - \text{Taxa de saída}$$

determine a equação diferencial que modela a massa $y = y(t)$ (em Kg) de poluente no tanque no instante t.

(b) Para

$$y0 = 1\,\text{kg}, \quad c(t) = e^{-2t}\,\text{kg/m}^3,$$

determine uma aproximação da massa de poluente no tanque após $T = 4$ minutos pelos métodos de Euler, RK do ponto médio, RK de Euler modificado e RK de ordem 4, considerando 2, 4, 8 e 16 subdivisões do intervalo $[0, 4]$.

(c) Sabendo que a solução exata do problema é

$$y(t) = -\frac{100}{19}\left(e^{-2\,t} - 1.19\,e^{-0.1t}\right),$$

estude a evolução do erro com a diminuição do espaçamento para cada método numérico utilizado. Em particular, ilustre que o valor da ordem de convergência numérica verificada

$$p \approx \log_2\left(\frac{|e_h|}{|e_{h/2}|}\right)$$

parece convergir para os valores teóricos esperados em cada método.

Resolução.

(a) Seja $y = y(t)$ a massa de poluente no tanque no instante t. Então, a taxa de entrada de massa de poluente é dada por

$$\text{Taxa de entrada} = c(t) \times 10\,\text{kg/min}$$

enquanto que a taxa de saída é

$$\text{Taxa de saída} = \underbrace{\frac{y(t)}{100}}_{\text{Concentração}} \times 10\,\text{kg/min} = 0.1y(t)\,\text{kg/min}.$$

Assim, a equação diferencial que rege a massa de poluente no tanque é dada por

$$y'(t) = 10c(t) - 0.1y(t).$$

(b) Utilizando os algoritmos dos exercícios 9.10, 9.23, 9.29 e 9.34, temos a tabela seguinte com as aproximações para cada método e cada espaçamento.

h	Euler	RK P. méd.	RK Eul. mod.	RK ord. 4
2.00	17.0063127778	1.2648424108	7.5324679763	4.3707206919
1.00	9.2319440660	3.4398974787	5.1802824824	4.2108636334
0.50	6.3588469079	4.0103717929	4.4549593395	4.1975216684
0.25	5.1895807599	4.1509396260	4.2620592884	4.1966163810

(c) Temos os gráficos das aproximações sucessivas para os vários espaçamentos na figura 9.6.

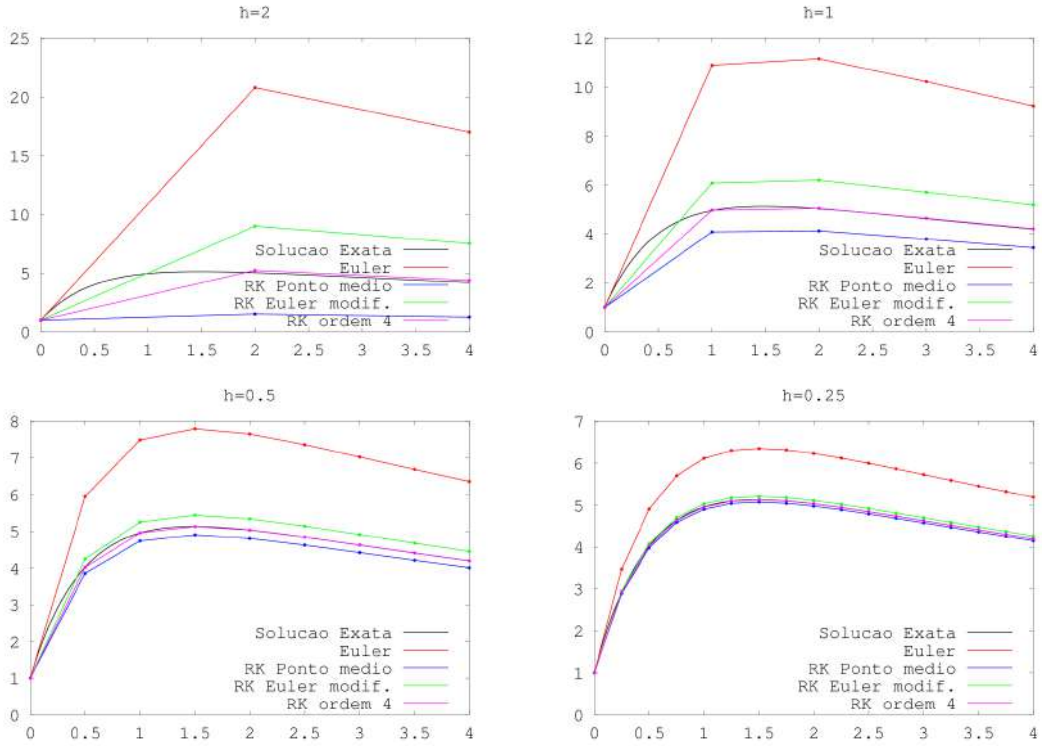


Figura 9.6: *Gráficos da soluções exata e aproximações pelos 4 métodos do exercício 9.36.*

Temos também os erros das aproximações e os valores da ordem de convergência numérica na tabela 9.2. Os resultados ilustram que à medida que h diminui e uma vez que a solução é infinitamente diferenciável, para um método com ordem de convergência p quando o espaçamento passa a metade o erro decresce aproximadamente com um fator de 2^{-p}. Nessa perspectiva, a melhor aproximação é obtida pelo método RK de ordem 4,

h	Euler		RK P. méd.		RK Eul. mod.		RK ord. 4	
	$\lvert y(T) - y_n \rvert$	p	$\lvert y(T) - y_n \rvert$	p	$\lvert y(T) - y_n \rvert$	p	$\lvert y(T) - y_n \rvert$	p
2.00	12.8098	1.3	2.93171	2.0	3.33591	1.8	0.174166	3.6
1.00	5.03539	1.2	0.756657	2.0	0.983728	1.9	0.0143089	3.9
0.50	2.16229	1.1	0.186183	2.0	0.258405	2.0	0.000966973	4.0
0.25	0.993026	-	0.0456151	-	0.0655046	-	6.16855e-005	-

Tabela 9.2: *Erro e ordem de convergência numérica do exercício 9.36.*

seguindo-se o par de métodos RK do ponto médio e de Euler modificado e, por último, o método de Euler.

Capítulo 10

Método das Diferenças Finitas

O método das diferenças finitas (MDF) será abordado neste capítulo como uma introdução à resolução numérica de equações diferenciais às derivadas parciais (EDP). Como exemplo ilustrativo vamos considerar a equação do Calor (que modela a transferência de calor), sendo que os esquemas de diferenças finitas podem ser aplicados a outras equações diferenciais parciais como a de Laplace (que modela estados de equilibrio), das Ondas (que modela a propagação das ondas) ou de Maxwell (que modela fenómenos electromagnéticos).

Os métodos numéricos de diferenças finitas partem do desenvolvimento em série de Taylor de uma função, permitindo obter aproximações das suas derivadas, utilizando nós igualmente espaçados com espaçamento h. Tomemos o desenvolvimento em série de Taylor de uma função $y : \mathbb{R} \to \mathbb{R}$ dado por

$$y(t+h) = y(t)+y'(t)h+y''(t)\frac{h^2}{2}+y'''(t)\frac{h^3}{6}+y^{(4)}(t)\frac{h^4}{24}+y^{(5)}(t)\frac{h^5}{120}+O(h^6). \quad (10.1)$$

Temos assim a aproximação de **primeira ordem para a primeira derivada** por diferenças finitas progressiva em t, dada por

$$y'(t) \approx \frac{y(t+h)-y(t)}{h}. \quad (10.2)$$

A aproximação diz-se de primeira ordem, porque o erro é da ordem de h, uma vez que

$$y'(t) = \frac{y(t+h)-y(t)}{h} - y''(t)\frac{h}{2} + O(h^2),$$

e diz-se progressiva porque utiliza o instante $t+h$ para obter uma aproximação da derivada no instante t. Note-se que esta diferença finita é a base para o método de Euler, conforme descrito na secção 9.1.

Por outro lado, considerando o desenvolvimento de Taylor

$$y(t-h) = y(t) - y'(t)h + y''(t)\frac{h^2}{2} - y'''(t)\frac{h^3}{6} + y^{(4)}(t)\frac{h^4}{24} - y^{(5)}(t)\frac{h^5}{120} + O(h^6). \quad (10.3)$$

temos a aproximação de **primeira ordem para a primeira derivada** por diferenças finitas regressiva em t dado por

$$y'(t) \approx \frac{y(t) - y(t-h)}{h}. \quad (10.4)$$

Subtraindo (10.3) à equação (10.1), temos

$$y'(t) = \frac{y(t+h) - y(t-h)}{2h} + y'''(t)\frac{h^2}{3} + O(h^4)$$

e logo a aproximação de **segunda ordem para a primeira derivada** centrada em t dada por

$$y'(t) \approx \frac{y(t+h) - y(t-h)}{2h}. \quad (10.5)$$

A aproximação diz-se de segunda ordem, porque o erro decresce com h^2 e denomina-se centrada em t uma vez que utiliza os nós $t+h$ e $t-h$ em torno de t.

Quanto à segunda derivada, somando (10.1) e (10.3), temos

$$y''(t) = \frac{y(t+h) - 2y(t) + y(t-h)}{h^2} + y^{(4)}(t)\frac{h^2}{12} + O(h^6)$$

logo obtemos aproximação de **segunda ordem para a segunda derivada** centrada em t dada por

$$y''(t) \approx \frac{y(t+h) - 2y(t) + y(t-h)}{h^2}. \quad (10.6)$$

Passamos agora considerar as aproximações anteriores num exemplo concreto de modelação de fenómenos físicos por equações diferenciais parciais: A equação do calor. Por simplicidade, vamos considerar uma dimensão espacial e uma difusão de calor homogénea no meio. Como já foi referido no ínicio da secção, princípios semelhantes aos que vamos estudar podem ser considerados para outras equações, como por exemplo a equação de Laplace (que modela estados de equilíbrio) ou a equação das Ondas (que modela a propagação de ondas).

10.1 Equação do Calor

A equação do calor dada por

$$\frac{\partial u}{\partial t} = \mathrm{div}(D\nabla u) \tag{10.7}$$

modela a difusão de calor no espaço ao longo do tempo t, em que $u = u(x,t)$ é a temperatura no ponto $x \in \mathbb{R}^3$ e no instante $t \geq 0$, $D = D(x,t) \geq 0$ é o coeficiente de difusão, ∇ é o operador gradiente dado por

$$\nabla u = \left[\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial u}{\partial x_3}\right]$$

e o operador divergência é dado por

$$\mathrm{div} u = \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} + \frac{\partial u}{\partial x_3}.$$

A equação do calor (10.7) pode ser escrita como

$$\frac{\partial u}{\partial t} = \nabla D \cdot \nabla u + D\Delta u \tag{10.8}$$

em que o laplaciano Δ é dado por

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}.$$

Assim, num meio homogéneo em que o coeficiente de difusão D é constante, temos

$$\frac{\partial u}{\partial t} = D\Delta u. \tag{10.9}$$

Por uma questão de simplicidade trataremos apenas o caso de meio homogéneo (i.e. $D \geq 0$ constante) neste texto.

10.2 Caso unidimensional

Consideremos então uma barra metálica (unidimensional) com coeficiente de difusão D constante. A difusão de calor na barra é modelada pela EDP

$$\frac{\partial u}{\partial t}(x,t) = D\frac{\partial^2 u}{\partial x^2}(x,t), \quad x \in [a,b], \quad t \geq 0. \tag{10.10}$$

Para que o problema esteja bem posto (i.e., exista solução única) é necessária ainda a condição inicial (i.e., a distribuição de temperatura inicial da barra) dado por

$$u(x,0) = g(x), \quad x \in [a,b],$$

para uma função g conhecida e as condições fronteira (i.e., a temperatura nos dois extremos da barra ao longo do tempo) dadas por

$$u(a,t) = f_a(t), \quad u(b,t) = f_b(t), \quad t \geq 0.$$

Para obter a distribuição de temperaturas no instante T, temos de resolver o problema de valores fronteira (PVF)

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = D\dfrac{\partial^2 u}{\partial x^2}(x,t), & x \in [a,b], \quad t \geq 0, \\ u(x,0) = g(x), & x \in [a,b], \\ u(a,t) = f_a(t), & t \geq 0, \\ u(b,t) = f_b(t), & t \geq 0. \end{cases} \tag{10.11}$$

Pretendemos ilustrar alternativas para resolver este problema numericamente pelo método das diferenças finitas. Começamos por considerar a discretização

$$x_0 = a,\ x_1 = x_0 + h_x,\ x_2 = x_1 + h_x, \ldots, x_n = x_0 + nh = b \tag{10.12}$$

$$t_0 = 0,\ t_1 = h_t,\ t_2 = 2h_t, \ldots, t_m = mh_t = T \tag{10.13}$$

com espaçamento espacial $h_x = (b-a)/n$ e passo temporal $h_x = T/m$. Por facilidade de notação, consideremos

$$u_i^k = u(x_i, t_k), \quad , k = 0, 1, \ldots, m, \quad i = 0, 1, \ldots, n. \tag{10.14}$$

Notamos que as condições iniciais e de fronteira se traduzem agora em

$$u_i^0 = g(x_i), i = 0, 1, \ldots, n; \tag{10.15}$$

$$u_0^k = f_a(t_k), \quad u_n^k = f_b(t_k), k = 0, 1, \ldots, m \tag{10.16}$$

Assim, pela aproximação (10.2) temos

$$\frac{\partial u}{\partial t}(x_i, t_k) = \frac{u(x_i, t_{k+1}) - u(x_i, t_k)}{h_t} + O(h_t) \tag{10.17}$$

ou seja, a aproximação de primeira ordem para a primeira derivada em t dada por

$$\frac{\partial u}{\partial t}(x_i, t_k) \approx \frac{u_i^{k+1} - u_i^k}{h_t}.$$

Por outro lado, pela aproximação (10.6) temos

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_k) = \frac{u(x_{i+1}, t_k) - 2u(x_i, t_k) + u(x_{i-1}, t_k)}{h_x^2} + O(h_x^2). \tag{10.18}$$

ou seja, a aproximação de segunda ordem ordem para a primeira derivada em t dada por

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_k) \approx \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2}$$

Estas aproximações serão a base para obter vários métodos de diferenças finitas, como ilustramos de seguida.

10.2.1 Método Explícito

Começamos pelo caso mais simples. Substituindo as aproximações que derivam de (10.17) e (10.18) na equação (10.9) temos

$$\frac{u_i^{k+1} - u_i^k}{h_t} = D\frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2}, \tag{10.19}$$

de onde sai o esquema de diferenças finitas de ordem 1 em h_t e de ordem 2 em h_x progressivo no tempo e centrado no espaço (PTCE[1]). Este método é também chamado de método explícito, pois para aproximar os valores de u no instante t_{k+1} apenas utiliza os valores de u no instante anterior t_k.

Definição 10.1 (Esquema explícito para Eq. Calor).

Chama-se **esquema PTCE** ou **método explicito centrado no tempo** para o problema (10.11) referente à equação do calor num meio homogéneo ao esquema de diferenças finitas

$$\begin{cases} u_i^0 = g(x_i), \quad i = 0, 1, \ldots, n \\ u_0^k = f_a(t_k), \quad k = 0, 1, \ldots, m \\ u_n^k = f_b(t_k), \quad k = 0, 1, \ldots, m \\ u_i^{k+1} = u_i^k + D\dfrac{h_t}{h_x^2}\left(u_{i+1}^k - 2u_i^k + u_{i-1}^k\right),\ i = 1, \ldots, n-1,\ k = 0, \ldots, m-1 \end{cases} \tag{10.20}$$

[1]Geralmente este esquema é conhecido na literatura como FTCS - forward in time, centered in space.

A definição anterior levanta a questão da convergência do método (10.20) e em que circunstâncias isso acontece. Para estabelecer esse resultado, precisamos das três definições seguintes sobre métodos numéricos.

Definição 10.2 (Consistência).
Um método numérico diz-se **consistente de ordem** p se o erro de discretização for $O(h^p)$.

A consistência mede a discretização local do método. Em particular um método é consistente de ordem p se ao substituir os valores no instante corrente pelos valores exatos, o erro cometido for de ordem $O(h^p)$. Nos métodos que aqui apresentamos, a consistência é garantida pelo desenvolvimento em série de Taylor. Assim para soluções suficientemente suaves, as expressões (10.17) e (10.18) garantem que o método PTCE (10.20) é consistente com ordem 1 no tempo e 2 no espaço.

Note-se que consistência não implica diretamente convergência, pois não considera que os erros no instante t_k das aproximações obtidas através das iterações anteriores. A consistência apenas está relacionada com a discretização e não com a possível propagação de erros pelas aproximações sucessivas nos vários instantes de tempo anteriores. Recordemos então a definição de convergência neste contexto.

Definição 10.3 (Convergência).
Um método numérico diz-se **convergente de ordem** p se o erro cometido for da ordem de $O(h^p)$, isto é, se

$$|Y(T, x_i) - y_i^m| = O(h^p).$$

Assim, entre a consistência e a convergência existe uma diferença, que está então relacionada com a possível propagação de erros pelas iterações. Temos então a seguinte definição.

Definição 10.4 (Estabilidade).
Um método numérico diz-se **estável** (ou **bem condicionado**) se a pequenos erros nos dados corresponderem pequenos erros na aproximação final, isto é, se o erro não for amplificado nos vários passos do método.

Em certas condições, a estabilidade representa o complemento da consistência para garantir convergência. Temos então o seguinte resultado.

Teorema 10.5 (Teorema de equivalência de Lax[2]).

Consideremos um método numérico de diferenças finitas consistente e linear (na solução). Então o método é convergente (com a mesma ordem de consistência) se e só se é estável.

Demonstração. Não faremos aqui a prova. No entanto esta segue sensivelmente os mesmos passos das provas dos teoremas 9.5 e 9.19, em que a ordem de consistência e a estabilidade são usadas para garantir convergência da mesma ordem dos respetivos métodos. □

O resultado anterior estabelece a equivalência entre estabilidade e consistência, no caso de EDP lineares na sua solução, como é o caso da equação do calor. Assim, para mostrar convergência do método explícito (10.20) basta mostrar em que condições é estável.

Teorema 10.6.

O método explícito (10.20) *é estável se e só se*

$$D\frac{h_t}{h_x^2} \leq \frac{1}{2}. \tag{10.21}$$

Demonstração. Suponhamos que os elementos u_i^n estão afetados de erros e_i^n, ou seja, em vez de utilizarmos os valores exatos u_i^n, vamos considerar os valores afetados de erros

$$\tilde{u}_i^n = u_i^n + e_i^n.$$

Assim, temos que a propagação de erros é dada por

$$\begin{aligned} e_i^{k+1} =& \tilde{u}_i^{k+1} - u_i^{k+1} \\ =& \tilde{u}_i^k + D\frac{h_t}{h_x^2}\left(\tilde{u}_{i+1}^k - 2\tilde{u}_i^k + \tilde{u}_{i-1}^k\right) - u_i^k - D\frac{h_t}{h_x^2}\left(u_{i+1}^k - 2u_i^k + u_{i-1}^k\right) \\ =& \left(1 - 2D\frac{h_t}{h_x^2}\right) e_i^k + D\frac{h_t}{h_x^2} e_{i+1}^k + D\frac{h_t}{h_x^2} e_{i-1}^k \end{aligned}$$

Assim, com a condição (10.21) temos

$$1 - 2D\frac{h_t}{h_x^2} \geq 0$$

[2]Peter Lax foi um dos poucos matemáticos distinguido com o prémio Abel em 2005. Este prémio na área de Matemática distingue matemáticos brilhantes e com contribuições fundamentais e é atribuído pelo Rei da Noruega. Lax foi o terceiro a ter essa honra, uma vez que o prémio apenas começou a ser galardoado em 2003.

logo pela desigualdade de triangular, temos

$$\begin{aligned}|e_i^{k+1}| &\leq \left(1 - 2D\frac{h_t}{h_x^2}\right)|e_i^k| + D\frac{h_t}{h_x^2}(|e_{i+1}^k| + |e_{i-1}^k|) \\ &\leq \left(1 - 2D\frac{h_t}{h_x^2} + 2D\frac{h_t}{h_x^2}\right)\max_{i=0,1,\ldots,n}|e_i^k| \\ &\leq \max_{i=0,1,\ldots,n}|e_i^k|,\end{aligned}$$

ou seja, a propagação do erro para instante t_{n+1} é controlada (menor ou igual do que o máximo do erro no instante t_n. Note-se que no caso de $i = 1$ ou $i = n - 1$, os erros e_0 e e_n, respetivamente, são nulos (são valores fronteira), não afetando a demonstração.

Se por outro lado a condição (10.21) não for satisfeita, temos

$$2D\frac{h_t}{h_x^2} - 1 > 0 \Rightarrow 2D\frac{h_t}{h_x^2} > 1,$$

logo no caso particular $e_{i-1}^k = e_{i+1}^k = -e_i^k$, temos

$$\begin{aligned}e_i^{k+1} &= \left(1 - 2D\frac{h_t}{h_x^2}\right)e_i^k - 2D\frac{h_t}{h_x^2}e_i^k \\ &= \left(1 - 4D\frac{h_t}{h_x^2}\right)e_i^k\end{aligned}$$

logo

$$|e_i^{k+1}| = \left(4D\frac{h_t}{h_x^2} - 1\right)|e_i^k| > 2D\frac{h_t}{h_x^2}|e_i^k| > |e_i^k|,$$

ou seja, o erro no instante t_n é amplificado para o instante t_{n+1}, pelo que o resultado está demonstrado. □

Temos então o seguinte corolário, que sai diretamente do teorema de Lax 10.5.

Corolário 10.7 (Convergência do método explícito).

O esquema PTCE ou método explícito centrado no tempo (10.20) *para a equação do calor é convergente com ordem 1 em h_t e 2 em h_x se e só se a condição de estabilidade* (10.21) *for satisfeita.*

O resultado anterior mostra que devemos decrescer os valores dos passo temporal h_t e espaçamento espacial h_x tornando os tão pequenos quanto possível, mas mantendo a relação

$$\frac{h_t}{h_x^2} \approx C \leq \frac{1}{2D}.$$

Em particular isto mostra que o passo em t tem de decrescer muito mais rapidamente, aumentando por isso o tempo de cálculo.

Exercício Octave 10.8. Escreva uma função em Octave

`ExplicitoCalor(expg,expfa,expfb,D,a,b,T,hx,ht)`

que resolva numericamente o problema (10.11) com coeficiente de difusão D constante e `expg`, `expfa` e `expfb` as expressões das funções g, f_a e f_b, respetivamente, com variáveis t e x pelo método explícito (10.20) com passo temporal `ht` e espaçamento espacial `hx`, de forma a obter uma aproximação da solução $u(x,t)$, $x \in [a,b]$, $t \in [0,T]$. O algoritmo deve verificar se a condição de estabilidade (10.21) é verificada.

Resposta.

No ficheiro `ExplicitoCalor.m` escrevemos:

```
function u=ExplicitoCalor(expg,expfa,expfb,D,a,b,T,hx,ht)
if D*ht/hx^2 <= 0.5
  disp('Condicao de convergencia satisfeita!')
else
  disp('Condicao de convergencia NAO satisfeita!')
end
nx=round((b-a)/hx);
nt=round(T/ht);
lx=a:hx:b;
lt=0:ht:T;
u=zeros(nt+1,nx+1);
g= vectorize(inline(expg,'x'));
fa= vectorize(inline(expfa,'t'));
fb= vectorize(inline(expfb,'t'));
%condicao inicial
u(1,:)=g(lx);
%condicao fronteira
u(:,1)=fa(lt);
u(:,end)=fb(lt);
%Iteracoes
for kk=1:nt
  for ii=2:nx
    u(kk+1,ii)=u(kk,ii)+D*ht/hx^2*...
       (u(kk,ii+1)-2*u(kk,ii)+u(kk,ii-1));
  end
```

```
end
figure(1)
plot(lx,u(end,:),'r-');
title(['Aproximacao no instante t=',num2str(nt*ht)])
xlabel('x');
ylabel('Temperatura');
figure(2)
[t,x]=meshgrid(lt,lx);
surf(x,t,u.','EdgeColor','none','FaceColor','interp');
title(['Aproximacao ao longo do tempo']);
xlabel('x');
ylabel('Tempo (t)');
zlabel('Temperatura');
return
```

Note-se que no algoritmo anterior, o ciclo interior

```
for ii=2:nx
  u(kk+1,ii)=u(kk,ii)+D*ht/hx^2*...
     (u(kk,ii+1)-2*u(kk,ii)+u(kk,ii-1));
end
```

pode ser substituído pela linha seguinte, para diminuir o tempo de cálculo:

```
u(kk+1,2:end-1)=u(kk,2:end-1)+...
  D*ht/hx^2*(u(kk,3:end)-2*u(kk,2:end-1)+u(kk,1:end-2));
```

Exercício 10.9. Considere o problema

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = \dfrac{1}{2}\dfrac{\partial^2 u}{\partial x^2}(x,t), & x \in [-1,1], \quad t \in [0,1] \\ u(x,0) = \sin(\pi x), & x \in [-1,1], \\ u(1,t) = u(-1,t) = 0, & t \in [0,1]. \end{cases} \tag{10.22}$$

que modela a temperatura de uma barra compreendida no intervalo $[-1,1]$ com temperatura inicial $u(x,0) = \sin(\pi x)$ e cujas extremidades são mantidas a temperatura nula ao longo do tempo. A solução exata do problema é dada por

$$u(x,t) = e^{-\pi^2 t/2}\sin(\pi x).$$

(a) Determine uma aproximação da solução pelo método explicito (10.20) com espaçamento espacial e passo temporal dados por

(i) $h_x = 0.1$ e $h_t = h_x^2 = 0.01$;

(ii) $h_x = 0.05$ e $h_t = h_x^2 = 0.0025$;

(iii) $h_x = 0.025$ e $h_t = h_x^2 = 0.000625$;

Compare o erro cometido no instante T, isto é,

$$e_{h_t,h_x} = \max_{i=1,2,\ldots,n} |u(T, x_i) - u_i^m|.$$

(b) Determine uma aproximação da solução pelo método explicito (10.20) com espaçamento espacial $h_x = 0.05$ e passo temporal $ht = 0.01$. O que conclui?

Resposta.

(a) Para o problema dado, temos

$$a = -1, \quad b = 1, \quad T = 1, \quad D = \frac{1}{2},$$

a condição inicial definida pela função $g(x) = \sin(\pi x)$ e as condições fronteira definidas pelas funções $f_a(t) = f_b(t) = 0$. Notamos também que dado o valor de D, a condição de estabilidade (10.21) é satisfeita se $h_t/h_x^2 \leq 1$, o que acontece em todos os casos nesta alínea. Através da função Octave do exercício 10.8 obtemos as aproximações da figura 10.1 Temos também que o máximo do erro é dado por

	e_{h_t,h_x}
i)	5.7532×10^{-4}
ii)	1.4543×10^{-4}
iii)	3.6455×10^{-5}

O resultado vai de encontro ao esperado. Como o passo em t diminui por um fator de $1/4$ e o método explícito (10.20) tem ordem 1 em t, espera-se que o erro decresça por um fator de $1/4$, se o espaçamento h_x diminuir para garantir a mesma ordem. Assim, como o espaçamento em x diminui por um fator de $1/2$ e o método explícito (10.20) tem ordem 2 em x, espera-se que o erro global decresça por um fator de $1/4$, o que acontece de (i) para (ii) e depois de (ii) para (iii).

(b) Neste caso a condição de estabilidade (10.21) não é verificada, o método explicito (10.20) não é estável (e logo não é convergente), pelo que a solução encontrada na figura 10.2 não tem significado.

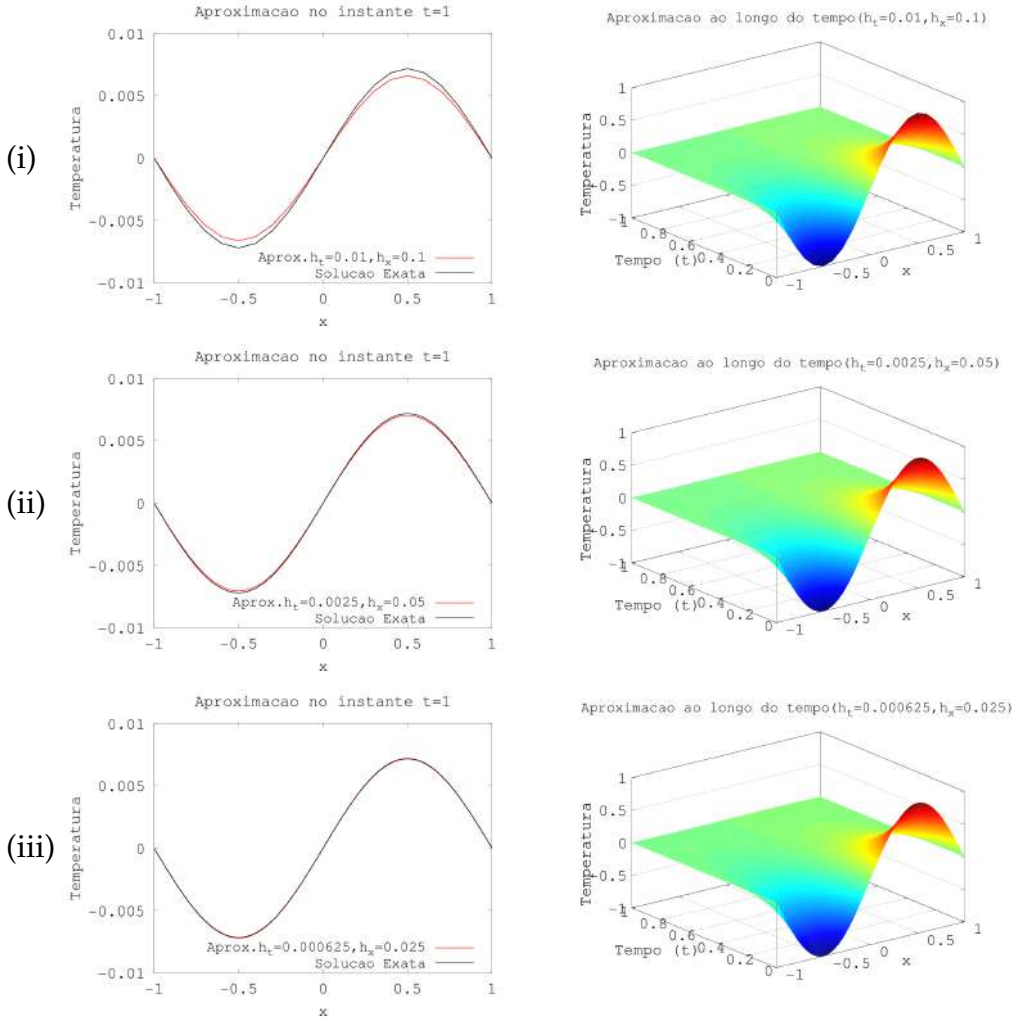


Figura 10.1: *Aproximações das soluções do exercício 10.9 (a).*

O método explícito (10.20) tem uma grande desvantagem: a condição de estabilidade 10.21 determina que h_t tenha de ser muito pequeno em relação a h_x, o que pode levar a tempos de cálculo muito elevados. Desta forma, pretende-se obter um método sem condição de estabilidade, o que faremos de seguida.

10.2.2 Método Implícito

Em vez de tomarmos a aproximação progressiva para a derivada no tempo, vamos considerar a aproximação regressiva (10.4), que por (10.3) é também de or-

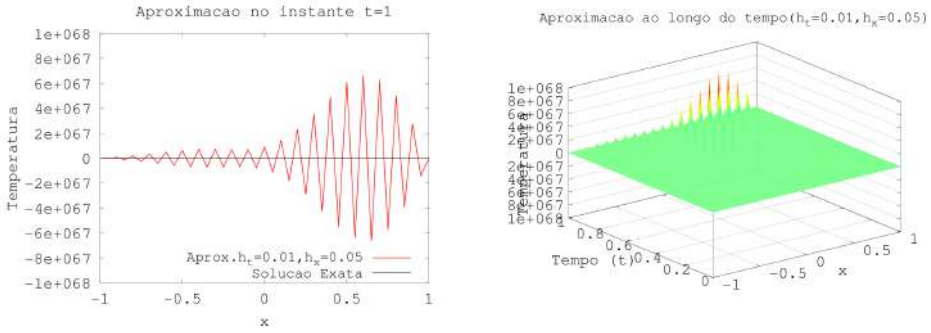


Figura 10.2: *Aproximações das soluções do exercício 10.9 (b).*

dem um no passo h_t. Desta forma, temos

$$\frac{\partial u}{\partial t}(x_i, t_k) = \frac{u_i^k - u_i^{k-1}}{h_t} + O(h_t) \tag{10.23}$$

e logo obtemos a aproximação de primeira ordem para a primeira derivada em t dada por

$$\frac{\partial u}{\partial t}(x_i, t_k) \approx \frac{u_i^k - u_i^{k-1}}{h_t}.$$

Assim, aplicando o mesmo princípio que para o método explicito (ou seja, tomando a aproximação (10.18) centrada de segunda ordem para a segunda derivada em x), substituindo na equação (10.9) temos o método implícito ou regressivo no tempo e centrado no espaço (RTCE)

$$\frac{u_i^k - u_i^{k-1}}{h_t} = D\frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2}.$$

Definição 10.10 (Esquema implícito para Eq. Calor).

Chama-se **esquema RTCE** ou **método implícito centrado no tempo** para o problema (10.11) referente à equação do calor num meio homogéneo ao esquema de diferenças finitas

$$\begin{cases} u_i^0 = g(x_i), \quad i = 0, 1, \ldots, n \\ u_0^k = f_a(t_k), \quad k = 0, 1, \ldots, m \\ u_n^k = f_b(t_k), \quad k = 0, 1, \ldots, m \\ \left(1 + 2D\dfrac{h_t}{h_x^2}\right) u_i^k - D\dfrac{h_t}{h_x^2}\left(u_{i+1}^k + u_{i-1}^k\right) = u_i^{k-1},\ i = 1, \ldots, n-1,\ k = 1, \ldots, m \end{cases} \tag{10.24}$$

Como veremos o método implícito é incondicionalmente estável, o que é a sua grande vantagem. No entanto, este denomina-se implícito, pois a cada instante t_k temos de resolver um sistema linear. Note-se que o valor de u_i^k depende dos valores de $u_{i\pm1}^k$ que não são conhecidos, pelo que se impõe a determinação de todos estes valores em conjunto pela resolução de um sistema linear.

Suponhamos então que conhecemos os valores u_i^{k-1} no instante t_{k-1} para determinado k inteiro, com $k \leq 1$. Assim de (10.24) temos a série de equações

$$\left(1 + 2D\frac{h_t}{h_x^2}\right) u_i^k - D\frac{h_t}{h_x^2}\left(u_{i+1}^k + u_{i-1}^k\right) = u_i^{k-1}, i = 1, \ldots, n-1,$$

com incógnitas $u_i^k, i = 1, \ldots, n-1$, uma vez que u_0^k e u_n^k são definidos pelas condições fronteira. Assim, para $i = 1$ temos

$$\left(1 + 2D\frac{h_t}{h_x^2}\right) u_1^k - D\frac{h_t}{h_x^2}u_2^k = u_i^{k-1} + D\frac{h_t}{h_x^2}\underbrace{u_0^k}_{f_a(t_k)}$$

e para $i = n-1$, temos

$$\left(1 + 2D\frac{h_t}{h_x^2}\right) u_{n-1}^k - D\frac{h_t}{h_x^2}u_{n-2}^k = u_i^{k-1} + D\frac{h_t}{h_x^2}\underbrace{u_n^k}_{f_b(t_k)}.$$

Assim é fácil ver que para cada instante k obtemos o sistema linear $A_k x_k = b_k$ em que a matriz tridiagonal A é dada por

$$A_k = \begin{bmatrix} 1 + 2D\frac{h_t}{h_x^2} & -D\frac{h_t}{h_x^2} & 0 & \ldots & 0 \\ -D\frac{h_t}{h_x^2} & 1 + 2D\frac{h_t}{h_x^2} & -D\frac{h_t}{h_x^2} & \ddots & \vdots \\ 0 & -D\frac{h_t}{h_x^2} & 1 + 2D\frac{h_t}{h_x^2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -D\frac{h_t}{h_x^2} \\ 0 & \ldots & 0 & -D\frac{h_t}{h_x^2} & 1 + 2D\frac{h_t}{h_x^2} \end{bmatrix} \tag{10.25}$$

e o vetor solução x_k e o vetor do segundo membro b_k são dados por

$$x_k = \begin{bmatrix} u_1^k \\ u_2^k \\ \vdots \\ u_{n-2}^k \\ u_{n-1}^k \end{bmatrix}, \quad b_k = \begin{bmatrix} u_1^{k-1} + D\dfrac{h_t}{h_x^2} f_a(t_k) \\ u_2^{k-1} \\ \vdots \\ u_{n-2}^{k-1} \\ u_{n-1}^{k-1} + D\dfrac{h_t}{h_x^2} f_b(t_k) \end{bmatrix} \tag{10.26}$$

É fácil verificar que o método implícito (10.24) é consistente com ordem 1 no tempo e 2 no espaço, a partir de (10.23) e de (10.18), à semelhança do argumento para o método explicito. Pelo teorema de Lax 10.5 para mostrar convergência da mesma ordem, precisamos apenas de mostrar estabilidade do método.

Teorema 10.11.
O método implícito (10.24) *é incondicionalmente estável.*

Demonstração. Daremos apenas uma ideia da prova. Grosso modo, basta mostrar que o sistema $A_k x_k = b_k$ é bem condicionado para a matriz (10.25) e os vetores (10.26). Por outras palavras, basta verificar que a matriz A é de fácil inversão, ou seja, que os seus valores próprios estão longe de zero. Pelo Teorema de Gershgorin 4.17, temos que os valores próprios da matriz A_k estão todos na bola de centro em $1 + 2Dh_t/h_x^2$ com raio $2Dh_t/h_x^2$. Em particular, isto mostra que qualquer que sejam os valores dos passo h_t e espaçamento h_x a matriz A tem valores próprios todos com módulo maior que 1, pelo que a sua inversão é bem condicionada. □

Temos então seguinte corolário, que sai diretamente do Teorema de Lax 10.5

Corolário 10.12 (Convergência do método implícito).
O esquema RTCE ou método implícito centrado no tempo (10.24) *para a equação do calor é incondicionalmente convergente com ordem 1 em h_t e 2 em h_x.*

Nota 10.13. A escolha entre usar o método implícito (10.24) ou explícito (10.20) é sobretudo computacional, uma vez que a ordem de convergência é igual. O primeiro impõe a resolução de um sistema linear a cada iteração, mas não tem

restrições em termos de número de iterações. Na segunda, cada iteração tem um custo baixo (é apenas uma atualização de valores), mas o número de iterações necessárias pode ser muito grande devido a um passo temporal muito pequeno imposto pela condição de estabilidade (10.21).

Exercício Octave 10.14. Escreva uma função em Octave

```
ImplicitoCalor(expg,expfa,expfb,D,a,b,T,hx,ht)
```

que resolva numericamente o problema (10.11) com coeficiente de difusão D constante e `expg`, `expfa` e `expfb` as expressões das funções g, f_a e f_b, respetivamente, com variáveis t e x pelo método implícito (10.24) com passo temporal `ht` e espaçamento espacial `hx`, de forma a obter uma aproximação da solução $u(x,t)$, $x \in [a,b]$, $t \in [0,T]$.

Resposta.

No ficheiro `ImplicitoCalor.m` escrevemos:

```
function u=ImplicitoCalor(expg,expfa,expfb,D,a,b,T,hx,ht)
nx=round((b-a)/hx);
nt=round(T/ht);
lx=a:hx:b;
lt=0:ht:T;
u=zeros(nt+1,nx+1);
g= vectorize(inline(expg,'x'));
fa= vectorize(inline(expfa,'t'));
fb= vectorize(inline(expfb,'t'));
%condicao inicial
u(1,:)=g(lx);
%condicao fronteira
u(:,1)=fa(lt);
u(:,end)=fb(lt);
b=zeros(nx-1,1)
A=(1+2*D*ht/hx^2)*eye(nx-1);
for ii=1:nx-2
  A(ii,ii+1) = -D*ht/hx^2;
  A(ii+1,ii) = -D*ht/hx^2;
end
%Iteracoes
for kk=1:nt
  b=(u(kk,2:nx))';
  b(1)=b(1)+u(kk+1,1)*D*ht/hx^2;
```

```
  b(end)=b(end)+u(kk+1,end)*D*ht/hx^2;
  u(kk+1,2:nx)=(A\ b)';
end
figure(1)
plot(lx,u(end,:),'r-');
h=title(['Aproximacao no instante t=',num2str(nt*ht)]);
xlabel('x');
ylabel('Temperatura');
figure(2)
[t,x]=meshgrid(lt,lx);
surf(x,t,u.','EdgeColor','none','FaceColor','interp');
title('Aproximacao ao longo do tempo');
xlabel('x');
ylabel('Tempo (t)');
zlabel('Temperatura');
return
```

Exercício 10.15. Considere o problema

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = 0.1\dfrac{\partial^2 u}{\partial x^2}(x,t), & x \in [-1,1], \quad t \in [0,1] \\ u(x,0) = 0, & x \in [-1,1], \\ u(1,t) = 2 - 2e^{-4t}, & t \in [0,1], \\ u(-1,t) = 0, & t \in [0,1]. \end{cases} \tag{10.27}$$

que modela a temperatura de uma barra compreendida no intervalo $[-1,1]$ com temperatura inicial nula e em que uma extremidade é mantida a zero graus e a é aquecida ao longo do tempo com temperatura $u(1,t) = 2 - 2e^{-4t}$.

Considere o espaçamento $h_x = 0.1$.

(a) Determine uma aproximação da solução pelo método explicito (10.20), considerando um passo h_t que garanta convergência.

(b) Determine uma aproximação da solução pelo método implícito (10.24) com passo $h_t = 0.1$.

(c) Compare as duas soluções.

Resposta.

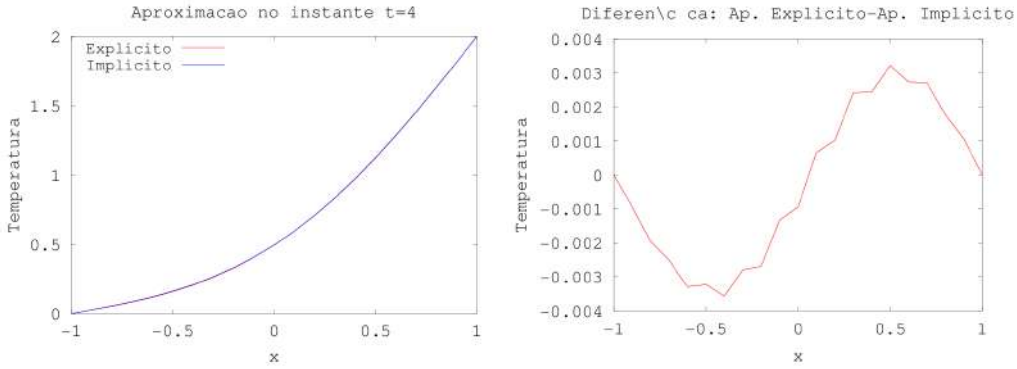


Figura 10.3: *Aproximações do exercício 10.15 para $t = 1$ e sua diferença.*

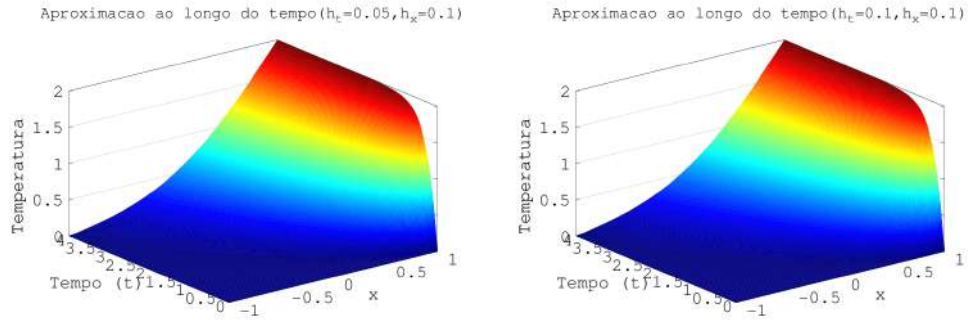


Figura 10.4: *Aproximações do método explícito (esquerda) e implícito (direita) do exercício 10.15 ao longo do tempo.*

(a) Temos neste caso

$$a = -1, \quad b = 1, \quad T = 1, \quad D = 0.1,$$

a condição inicial definida pela função $g(x) = 0$ e as condições fronteira definidas pelas funções $f_a(t) = 0$ e $f_b(t) = 2 - 2e^{-4t}$. Assim, começamos por verificar que para $h_x = 0.1$ de forma a garantir a condição de estabilidade (10.21) para o método explicito, temos

$$0.1\frac{h_t}{0.1^2} \leq \frac{1}{2} \quad \Rightarrow \quad h_t \leq 0.05.$$

Assim, escolhendo $h_t = 0.05$ temos a aproximação nas figuras 10.3 e 10.4, utilizando o algoritmo da exercício 10.8.

(b) Temos a aproximação nas figuras 10.3 e 10.4, utilizando o algoritmo da exercício 10.14.

(c) Temos a comparação na figura 10.3. De notar que as aproximações estão muito próximas, embora se espere que a do método explicito seja melhor pois usa um passo temporal menor (e ambos os métodos têm igual ordem de convergência).

O método implícito resolve um dos problemas do método explicito: é incondicionalmente estável. No entanto isto acarreta uma desvantagem, uma vez que a sua aplicação envolve a resolução de um sistema linear a cada iterada. Mais ainda, tanto um como outro método têm apenas convergência linear em h_t. Assim, a convergência quadrática em h_x é condicionada pela convergência (apenas) linear em h_t. Para se tirar partido da convergência quadrática, teríamos de diminuir os passo h_t e espaçamento h_x por fatores de $1/4$ e $1/2$, respetivamente. Note-se que esta relação no decrescimento entre h_t e h_x é semelhante à dada pela condição de estabilidade do método explícito (10.21), uma vez que nesse caso temos de garantir que $h_t/h_x^2 = C \leq 1/2$. Assim, embora se possa utilizar qualquer h_t em termos de convergência para o método implícito, em termos de aproveitar a ordem de convergência quadrática em x convém manter a relação $h_t/h_x^2 = C$.

Desta forma, vamos ver um último caso de método de diferenças finitas em que a convergência em t seja também quadrática.

10.2.3 Método de Crank-Nicolson

Para introduzir o método de Crank-Nicolson, vamos considerar o desenvolvimento em série de Taylor de y em torno de $t + \theta h$, com $0 \leq \theta \leq 1$. Assim, temos por um lado que

$$y(t+h) = y(t+\theta h) + (1-\theta)hy'(t+\theta h) + \frac{(1-\theta)^2h^2}{2}y''(t+\theta h) + O(h^3), \quad (10.28)$$

enquanto que por outro temos

$$y(t) = y(t+\theta h) - \theta hy'(t+\theta h) + \frac{\theta^2h^2}{2}y''(t+\theta h) + O(h^3). \quad (10.29)$$

Subtraindo as duas equações, temos

$$y(t+h) - y(t) = hy'(t+\theta h) + \frac{(1-2\theta)h^2}{2}y''(t+\theta h) + O(h^3),$$

uma vez que

$$(1-\theta)^2 - \theta^2 = 1 - 2\theta.$$

Notamos agora que para $\theta = 1/2$, o termo de ordem $O(h^2)$ se anula, pelo que obtemos nesse caso

$$y'\left(t+\frac{h}{2}\right) = \frac{y(t+h)-y(t)}{h} + O(h^2). \tag{10.30}$$

Analogamente temos que

$$\frac{\partial u}{\partial t}\left(x_i, t_k + \frac{h_t}{2}\right) = \frac{u_i^{k+1} - u_i^k}{h_t} + O(h_t^2) \tag{10.31}$$

e logo obtemos a aproximação de segunda ordem para a primeira derivada em t no ponto $(x_i, t_k + h/2)$dada por

$$\frac{\partial u}{\partial t}\left(x_i, t_k + \frac{h_t}{2}\right) \approx \frac{u_i^{k+1} - u_i^k}{h_t}. \tag{10.32}$$

Se por outro lado, se somarmos as equações (10.28) e (10.29), de novo para $\theta = 1/2$ obtemos que

$$y\left(t+\frac{h}{2}\right) = \frac{1}{2}\Big(y(t) + y(t+h)\Big) + O(h_t^2)$$

e analogamente, obtemos que

$$\frac{\partial^2 u}{\partial x^2}\left(x_i, t_k + \frac{h_t}{2}\right) = \frac{1}{2}\left[\frac{\partial^2 u}{\partial x^2}(x_i, t_k) + \frac{\partial^2 u}{\partial x^2}(x_i, t_{k+1})\right] + O(h_t^2).$$

Assim, tomando (10.18) no instante t_k e t_{k+1}, temos da equação anterior que

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}\left(x_i, t_k + \frac{h_t}{2}\right) =& \frac{1}{2}\left[\frac{u(x_{i+1}, t_k) - 2u(x_i, t_k) + u(x_{i-1}, t_k)}{h_x^2} + \right. \\ & \left. \frac{u(x_{i+1}, t_{k+1}) - 2u(x_i, t_{k+1}) + u(x_{i-1}, t_{k+1})}{h_x^2}\right] \\ & + O(h_t^2) + O(h_x^2).\end{aligned} \tag{10.33}$$

ou seja, temos a aproximação de ordem 2 no espaço e no tempo dada por

$$\frac{\partial^2 u}{\partial x^2}\left(x_i, t_k + \frac{h_t}{2}\right) \approx \frac{1}{2}\left[\frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2} + \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{h_x^2}\right] \tag{10.34}$$

Assim da equação do calor (10.9) no ponto $(x_i, t_h + h_t/2)$ e das aproximações (10.32) e (10.34), temos o método de Crank-Nicolson

$$\frac{u_i^k - u_i^{k-1}}{h_t} = \frac{D}{2}\left[\frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2} + \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{h_x^2}\right]$$

Nota 10.16. O método de Crank-Nicolson pode ser visto como uma média entre os métodos explícito e implícito, tendo a vantagem de com essa operação se melhorar a ordem de consistência. Analogamente, também o método RK de Euler modificado (9.19) pode ser visto da mesma forma em relação ao método de Euler (9.3).

Definição 10.17 (Esquema Crank-Nicolson para Eq. Calor).

Chama-se **método Crank-Nicolson** para o problema (10.11) referente à equação do calor num meio homogéneo ao esquema de diferenças finitas

$$\begin{cases} u_i^0 = g(x_i), \quad i = 0, 1, \ldots, n \\ u_0^k = f_a(t_k), \quad k = 0, 1, \ldots, m \\ u_n^k = f_b(t_k), \quad k = 0, 1, \ldots, m \\ \left(1 + \dfrac{Dh_t}{h_x^2}\right) u_i^{k+1} - \dfrac{Dh_t}{2h_x^2}\left(u_{i+1}^{k+1} + u_{i-1}^{k+1}\right) = \left(1 - \dfrac{Dh_t}{h_x^2}\right) u_i^k + \dfrac{Dh_t}{2h_x^2}\left(u_{i+1}^k + u_{i-1}^k\right), \\ \qquad\qquad i = 1, \ldots, n-1,\ k = 0, \ldots, m-1. \end{cases} \tag{10.35}$$

O método de Crank-Nicolson (10.35) envolve também a resolução de um sistema. Assim, temos o sistema $A_k x_k = b_k$ com

$$A_k = \begin{bmatrix} 1 + \dfrac{Dh_t}{h_x^2} & -\dfrac{Dh_t}{2h_x^2} & 0 & \ldots & 0 \\ -\dfrac{Dh_t}{2h_x^2} & 1 + \dfrac{Dh_t}{h_x^2} & -\dfrac{Dh_t}{2h_x^2} & \ddots & \vdots \\ 0 & -\dfrac{Dh_t}{2h_x^2} & 1 + \dfrac{Dh_t}{h_x^2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\dfrac{Dh_t}{2h_x^2} \\ 0 & \ldots & 0 & -\dfrac{Dh_t}{2h_x^2} & 1 + \dfrac{Dh_t}{h_x^2} \end{bmatrix}, \tag{10.36}$$

o vetor solução é $x_k = [u_1^k, u_2^k, \ldots, u_{n-1}^k]^T$ e o vetor do segundo membro b_k é dado

por

$$b_k = \begin{bmatrix} \left(1 - \dfrac{Dh_t}{h_x^2}\right) u_1^k + \dfrac{Dh_t}{2h_x^2} u_2^k + \dfrac{Dh_t}{2h_x^2} \left(f_a(t_{k+1}) + f_a(t_k)\right) \\ \left(1 - \dfrac{Dh_t}{h_x^2}\right) u_2^k + \dfrac{Dh_t}{2h_x^2} \left(u_1^k + u_3^k\right) \\ \vdots \\ \left(1 - \dfrac{Dh_t}{h_x^2}\right) u_{n-2}^k + \dfrac{Dh_t}{2h_x^2} \left(u_{n-3}^k + u_{n-1}^k\right) \\ \left(1 - \dfrac{Dh_t}{h_x^2}\right) u_{n-1}^k + \dfrac{Dh_t}{2h_x^2} \left(u_{n-2}^k\right) + \dfrac{Dh_t}{2h_x^2} \left(f_b(t_{k+1}) + f_b(t_k)\right) \end{bmatrix}. \tag{10.37}$$

No entanto, o método de Crank-Nicolson tem a grande vantagem de ter ordem quadrática tanto no tempo como no espaço.

Corolário 10.18 (Convergência do método Crank-Nicolson)*.*

O método de Crank-Nicolson (10.35) *para a equação do calor é incondicionalmente convergente com ordem quadrática tanto em h_t como em h_x.*

Demonstração. Por (10.31) e (10.33) temos que o método é consistente com ordem quadrática em h_t e em h_x. Assim, pelo teorema de Lax 10.5 basta mostrar estabilidade do método, de forma a obter o resultado. A prova de estabilidade é obtida por análise do condicionamento da matriz (10.36), à semelhança do que foi feito no teorema 10.11. □

Exercício 10.19. Considere o problema

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = 0.4 \dfrac{\partial^2 u}{\partial x^2}(x,t), & x \in [0,1], \quad t \in [0,1/2] \\ u(x,0) = \sin(\pi x), & x \in [0,1], \\ u(0,t) = u(1,t) = 0, & t \in [0,1/2]. \end{cases} \tag{10.38}$$

que modela a temperatura de uma barra compreendida no intervalo $[0,1]$ com temperatura inicial $g(x) = \sin(\pi x)$ e em que as extremidades são mantidas a zero graus.

(a) Verifique que a função

$$u(x,t) = e^{-0.4\pi^2 t} \sin(\pi x)$$

é solução do problema.

(b) Apresente as aproximações das soluções no instante final $t = 0.5$ pelo método Implícito (10.24) e pelo método de Crank-Nicolson (10.35) para passos e espaçamentos $h_t = h_x = 0.1$ e $h_t = h_x = 0.05$.

(c) Determine o erro máximo obtido em cada caso e estudo o seu decrescimento em função do espaçamento e passo.

Resposta.

(a) Temos

$$\frac{\partial u}{\partial t}(x,t) = -0.4\pi^2 e^{-0.4\pi^2 t}\sin(\pi x) = -0.4\pi^2 u(x,t)$$

e

$$\frac{\partial^2 u}{\partial x^2}(x,t) = -\pi^2 e^{-0.4\pi^2 t}\sin(\pi x) = -\pi^2 u(x,t)$$

logo a equação do calor é satisfeita com $D = 0.4$. Além disso temos a condição inicial

$$u(x,0) = e^0 \sin(\pi x) = \sin(\pi x)$$

e as condições de fronteira

$$u(0,t) = e^{-0.4\pi^2 t}\sin(0) = 0, \quad u(1,t) = e^{-0.4\pi^2 t}\sin(\pi) = 0,$$

logo a função u é solução do problema.

(b) Utilizando a função do exercício 10.14 e fazendo as alterações no sistema a resolver de acordo com (10.36) e (10.37) por forma a obter um algoritmo para o método de Crank-Nicolson (exercício), obtemos o gráficos das aproximações e respetivo erro da figura 10.5.

(c) Temos os máximos dos erros $\max_{i=0,1,\ldots,n} |u(0.5, x_i) - u_i^n|$ dados na tabela seguinte:

	Implícito	Crank-Nicolson
$h_x = h_t = 0.1$	0.052739	0.0013026
$h_x = h_t = 0.05$	0.026707	0.00032631

Conforme esperado (uma vez que a solução é infinitamente diferenciável), o erro decresce por uma fator de cerca de $1/4$ quando o espaçamento e passo passam a metade no método de Crank-Nicolson (ordem de convergência quadrática). No entanto, conforme esperado, no caso do método implícito o erro decresce apenas para metade, uma vez que a ordem de

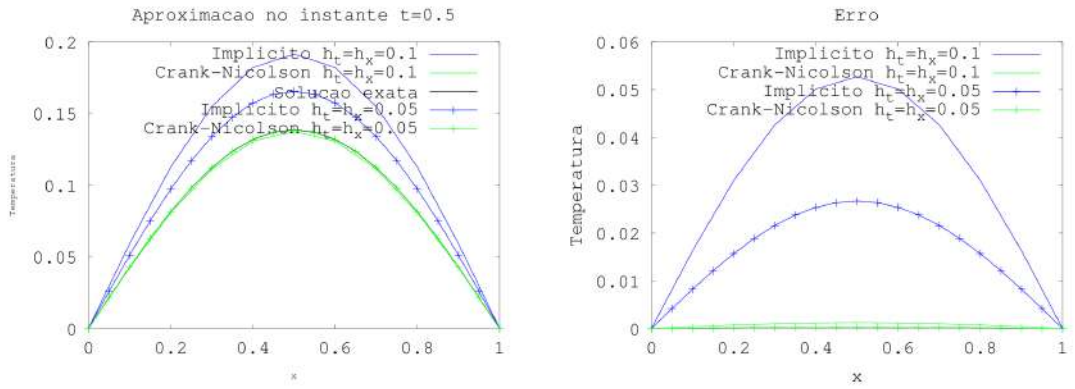


Figura 10.5: *Gráficos das aproximações no instante $t = 0.5$ (direita) e dos respetivos erros (esquerda) referentes ao exercício 10.19.*

convergência no tempo é apenas linear. Tal como se verifica, quando o espaçamento é pequeno os resultados por Crank-Nicolson são melhores que o do método implícito (ordem de convergência quadrática contra ordem de convergência apenas linear, respetivamente).

10.3 Caso Bidimensional

Consideramos agora o caso em que a coordenada espacial tem duas componentes, ou seja, $x = (x_1, x_2)$. A extensão para o caso tridimensional é em tudo semelhante, pelo que não a abordaremos neste texto.

Começamos então por considerar o domínio espacial

$$(x_1, x_2) \in [a_1, b_1] \times [a_2, b_2]$$

e mantemos o domínio temporal $t \in [0, T]$. Além da equação do calor com difusão homogénea (10.9), consideramos ainda a condição inicial

$$u(x_1, x_2, 0) = g(x_1, x_2), (x_1, x_2) \in [a_1, b_1] \times [a_2, b_2],$$

e as condições de fronteira de Dirichlet

$$\begin{aligned} &u(a_1, x_2, t) = f_1(x_2, t), \quad u(b_1, x_2, t) = f_2(x_2, t), \quad x_2 \in [a_2, b_2],\ t \in [0, T], \\ &u(x_1, a_2, t) = f_3(x_1, t), \quad u(x_1, b_2, t) = f_4(x_1, t), \quad x_1 \in [a_1, b_1],\ t \in [0, T]. \end{aligned} \tag{10.39}$$

Neste caso, temos que o laplaciano (espacial) de $u(x_1, x_2, t)$ é dado por

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$$

pelo que teremos de considerar a aproximação por diferenças finitas para cada uma das derivadas em x_1 e x_2. Consideramos a discretização espacial

$$x_{i,j} = (a_1 + ih_{x_1}, a_2 + jh_{x_2}), \quad i = 0, 1, \ldots, n_1,\ j = 0, 1, \ldots, n_2$$

com $n_1 + 1$ pontos na direcção x_1 e $n_2 + 1$ pontos na direcção x_2, igualmente espaçados em cada direcção por $h_{x_1} = \frac{b_1 - a_1}{n_1}$ e $h_{x_2} = \frac{b_2 - a_2}{n_2}$, respetivamente. mantemos a discretização temporal $t_k = kh_t$, para $k = 0, 1, \ldots, m$. Desta forma, considerando a notação $u_{i,j}^k := u(x_{i,j}, t_k)$, e considerando o esquema explicito, de forma semelhante ao que foi feito no caso unidimensional (10.19), temos a discretização por diferenças finitas para a equação do calor dada por

$$\frac{u_{i,j}^{k+1} - u_{i,j}^k}{h_t} = D\left(\frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{h_{x_1}^2} + \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j+1}^k}{h_{x_2}^2}\right), \tag{10.40}$$

que origina o esquema de diferenças finitas

$$u_{i,j}^{k+1} = u_{i,j}^k + D\frac{h_t}{h_{x_1}^2}\left(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k\right) + D\frac{h_t}{h_{x_2}^2}\left(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j+1}^k\right), \tag{10.41}$$

para $i = 1, 2, \ldots, n_1 - 1,\ j = 1, 2, \ldots, n_2 - 1,\ k = 0, 1, \ldots, m - 1$. Juntando os valores conhecidos das condições inicial e de fronteira, obtemos o método explicito centrado no tempo no caso bidimensional

$$\begin{cases} u_{i,j}^0 = g(x_{i,j}),, \quad i = 0, 1, \ldots, n_1,\ j = 0, 1, \ldots, n_2, \\ u_{0,j}^k = f_1(x_{2_j}, t_k),\ u_{n_1,j}^k = f_2(x_{2_j}, t_k),\ j = 0, 1, \ldots, n_2, k = 0, 1, \ldots, m \\ u_{i,0}^k = f_3(x_{1_i}, t_k),\ u_{i,n_2}^k = f_4(x_{1_i}, t_k),\ i = 0, 1, \ldots, n_1, k = 0, 1, \ldots, m \\ u_{i,j}^{k+1} = u_{i,j}^k + D\frac{h_t}{h_{x_1}^2}\left(u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k\right) + D\frac{h_t}{h_{x_2}^2}\left(u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j+1}^k\right), \\ \qquad\qquad i = 1, 2, \ldots, n_1 - 1,\ j = 1, 2, \ldots, n_2 - 1,\ k = 0, 1, \ldots, m - 1 \end{cases} \tag{10.42}$$

em que $x_{1_j} = a_1 + jh_{x_1}$ e $x_{2_j} = a_2 + jh_{x_2}$. A cada passo do método explicito, para se obter uma aproximação da solução $u_{i,j}^{k+1}$ em t_{k+1} utilizam-se os valores previamente obtidos de $u_{i,j}^k$, $u_{i\pm1,j}^k$ e $u_{i,j\pm1}^k$ em t_k, conforme a figura 10.6.

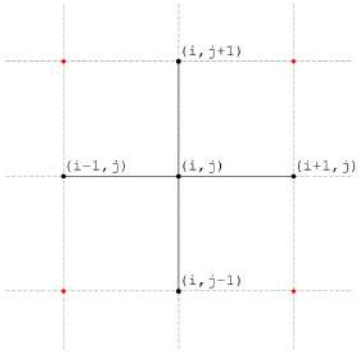


Figura 10.6: Nós utilizados em torno de (i, j) no esquema de diferenças finitas.

De notar que apenas se aplica o esquema para o nós interiores, uma vez que na fronteira temos a condição de fronteira de Dirichlet que define o valor da função. No caso dos esquemas implicitos e Crank-Nicholson é necessário resolver um sistema linear com $(n_1 - 1) \times (n_2 - 1)$ incógnitas, nomeadamente os valores de $u_{i,j}^{k+1}$ para $i = 1, 2, \ldots, n_1 - 1,\ j = 1, 2, \ldots, n_2 - 1$. Nesse sentido, é necessário renumerar os nós com a utilização apenas de um indice (em vez do duplo indice i, j) por forma a estabelecer esse sistema linear e fazer corresponder as incógnitas $u_{i,j}^{k+1}$ para $i = 1, 2, \ldots, n_1 - 1,\ j = 1, 2, \ldots, n_2 - 1$ aos $(n_1 - 1) \times (n_2 - 1)$ valores do vetor solução do sistema linear. A figura 10.7, exemplifica uma numeração possível para os casos de $n_1 = n_2 = 6, 8, 10$.

De notar que nos casos do método implícito e de Crank-Nicolson a matriz do sistema linear a resolver apenas terá no máximo 5 valores diferentes de zero em cada linha, nomeadamente os coeficientes de $u_{i,j}^{k+1}$, $u_{i\pm1,j}^{k+1}$ e $u_{i,j\pm1}^{k+1}$. Na realidade escolhendo uma numeração como a da figura 10.7, na linha ℓ da matriz, apenas os elementos das colunas $\ell, \ell-1, \ell+1, \ell+n_1-1$ e $\ell-n_1+1$ podem ser diferentes de zero. Isto torna a matriz do sistema apenas com 5 diagonias diferentes de zero, podenso-se aplicar métodos próprios para a resolução de sistemas desta forma. Note-se que em alguns casos de linhas ℓ alguns dos valores das colunas $\ell - 1, \ell + 1, \ell + n_1 - 1$ e $\ell - n_1 + 1$ podem ser zero, bastando para isso que exista um ou mais elementos vizinhos que estejam sobre a fronteira. Nesse ponto vizinho o valor de u deixa de ser incógnita para passar para o segundo membro do sistema linear com o valor dado pela condição de fronteira.

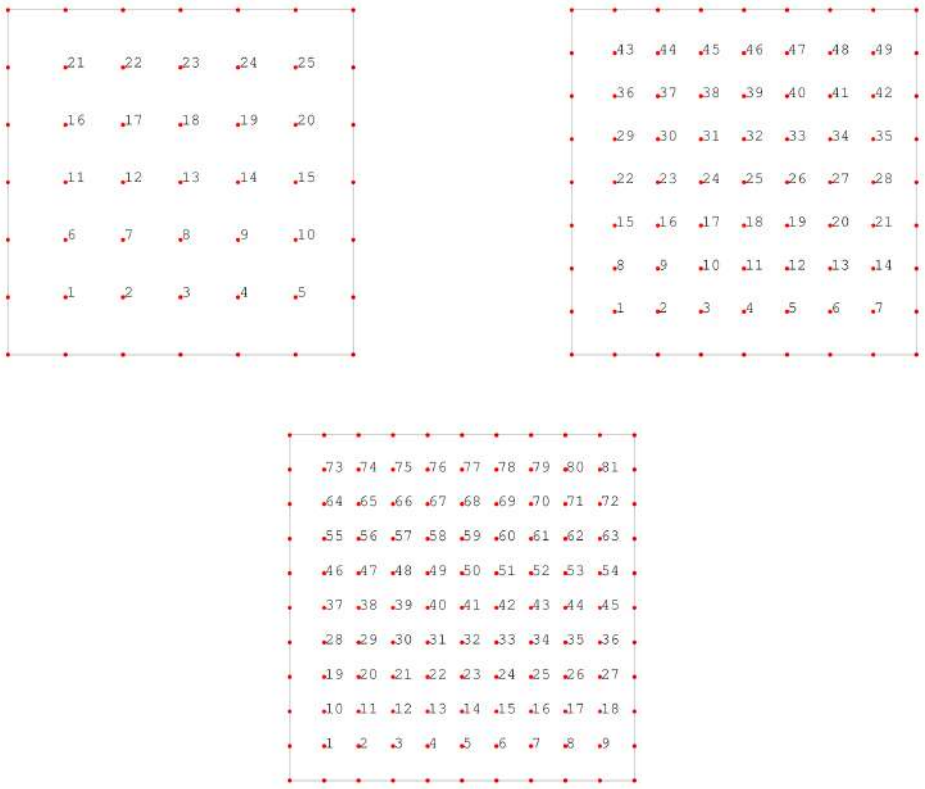


Figura 10.7: Possível numeração dos nós interiores para $n_1 = n_2 = 6, 8, 10$

.

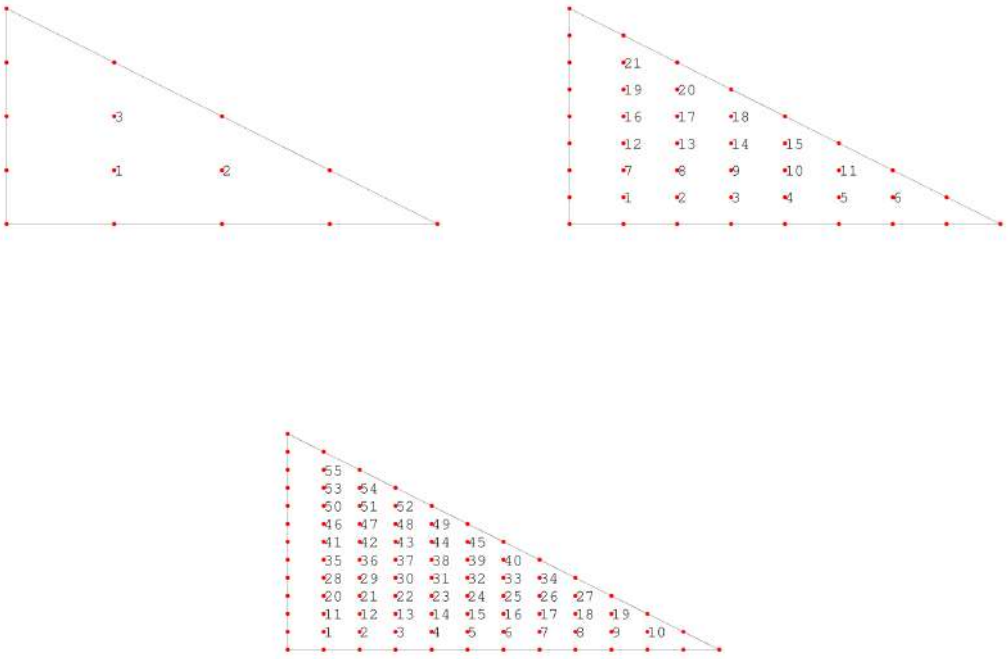


Figura 10.8: Possível numeração dos nós interiores para um triângulo retângulo com malhas com diferentes refinamentos.

10.3.1 Dominios poligonais

As diferenças finitas estão especialmente talhadas para domínios retangulares, uma vez que é necessário ter uma malha em que para cada ponto interior façam parte também da malha os pontos vizinhos segundo as direções cartesianas, devido ao esquema de diferenças finitas (ver figura 10.6). No entanto, sempre que seja possível manter esta caracteristica, os esquemas podem ser generalizados para domínios poligonais, conforme é indicado nas figuras 10.8 e 10.9, por exemplo.

As principais dificuldades para dominios poligonais são duas, que não serão abordadas neste texto:

- Para os esquemas implicito e Crank-Nicolson os sistema linear a resolver deixa de ter uma estrutura com 5 diagonais, pelo que os métodos para resolver este tipo de sistemas lineares não podem ser utilizados. A matriz

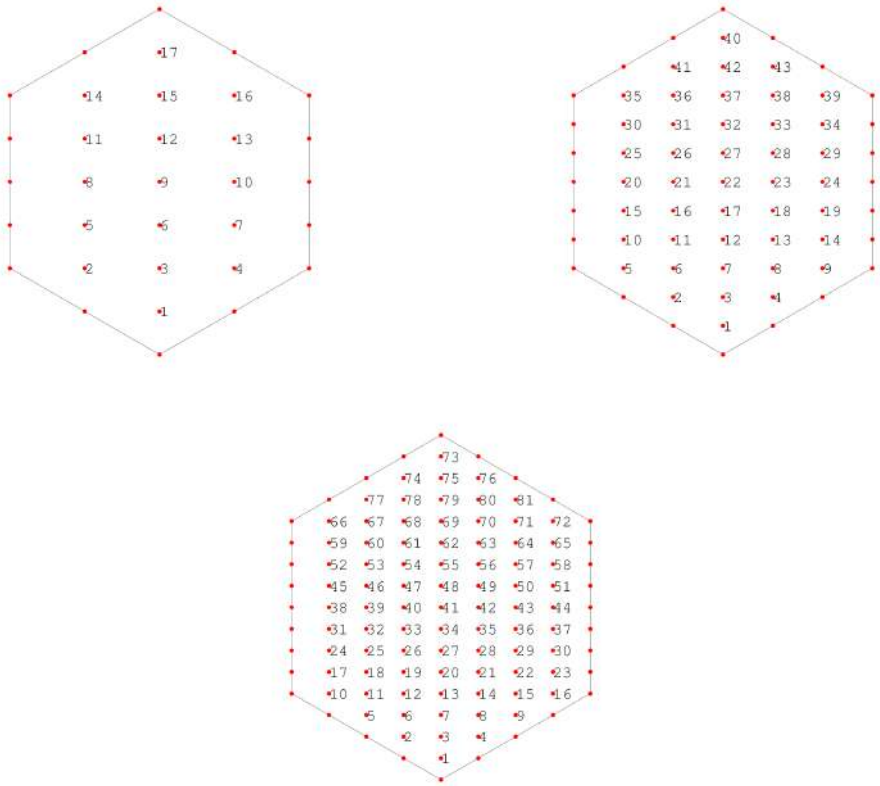


Figura 10.9: Possível numeração dos nós interiores para um hexágono regular com malhas com diferentes refinamentos.

mantém-se no entanto esparsa e com um máximo de 5 elementos não negativos por linha.

- Se em vez da condição de Dirichlet tivermos uma condição de Neumann na fronteira, isto é, uma condição para o valor da derivada normal $\frac{\partial u}{\partial \nu} = g$ sobre a fronteira, em que ν é o vetor unitário normal exterior à fronteira. Enquanto que no caso de dominios retangulares, como as fronteiras estão alinhados com os eixos cartesianos, os vetores normais à fronteira também estão e logo as derivadas normais podem ser reduzidas a derivadas na direção x_1 ou x_2, pelo que podemos aplicar esquemas de diferenças finitas para as aproximar em pontos sobre a malha (uma vez que os pontos da malha estão alinhados com a derivada normal). No caso de dominios poligonais, pode acontecer que o vetor normal num nó de fronteira não esteja alinhada com a malha, criando dificuldades adicionais para estabelecer um esquema de diferenças finitas com os nós existentes. Um exemplo é a derivada normal nos nós da hipotenusa do triângulo retângulo da figura 10.8.

Índice

Bibliografia

[1] M. Ramos. *Curso Elementar de Equações Diferenciais*. Universidade de Lisboa,2ª ed., 2000.

[2] Michael Renardy and Robert Rogers. *An introduction to partial differential equations*. Springer, 2^{nd} edition, 2003.

[3] George F. Simmons and Steven Krantz. *Equações diferenciais - Teoria, Técnica e Prática*. McGraw-Hill, 2008.

[4] Martin Schechter. *Principles of Functional Analysis*, volume 36 of *Graduate Studies in Mathematics*. American Mathemtics Society, 2^{nd} ed., 2001.

[5] Rainer Kress. *Numerical Analysis*. Springer, 1998.

[6] Luis T. Magalhães. *Álgebra linear como introdução a Matemática Aplicada*. Texto editora, 7ª ed., 1997.

[7] Jaime Campos Ferreira. *Introdução à Análise Real*. Fundação Calouste Gulbenkian, 1995.

[8] H. Anton. *Cálculo: um novo horizonte*. Bookman, 6ªEd., 1999.

[9] J. Hadamard. *Lectures on Cauchy's Problems in Linear Partial Differential Equations*. Dover, New York, 1952.

[10] David Kincaid and Ward Cheney. *Numerical analysis - Mathematics of Scientific Computing*, volume 2 of *Pure and Mathematics Undergraduate Texts*. American Mathemtics Society, 3^{rd} ed., 2009.

[11] Benjamin Fine and Gerhard Rosenber. *The Fundamental Theorem of Algebra*. Springer-Verlag, 1997.

[12] T. Apostol. *Cálculo*, volume Volume 1. Editorial Reverté, 1999.