Matemática Computacional

Adérito Luís Martins Araújo

Notas de apoio às aulas de Matemática Computacional do Mestrado Integrado em Engenharia Electrotécnica e de Computadores, no ano lectivo de 2017/2018.

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Figura da capa: Documento jurídico sumério YBC 3879, do período Ur III (c. 2100 a.C. – c. 2000 a.C.), onde é descrito um algoritmo geométrico para determinar as soluções de equações quadráticas. Fonte: `cdli.ucla.edu/pubs/cdlj/2009/cdlj2009_003.html`.

Conteúdo

Capítulo 1

Aritmética computacional

A análise numérica é a disciplina da matemática que se ocupa da elaboração e estudo de métodos que permitem obter, *de forma efectiva*, soluções numéricas para problemas matemáticos, quando, por uma qualquer razão, não podemos ou não desejamos usar métodos analíticos.

Para perceber melhor o que se pretende dizer por *de forma efectiva*, consideremos o problema do cálculo do determinante. Como é sabido, o determinante de uma matriz quadrada $A = (a_{ij})_{i,j=1}^{n}$ é dado pela expressão

$$\det(A) = \sum \pm a_{1i_1} \cdots a_{ni_n},$$

onde a soma é efectuada sobre todas as $n!$ permutações $(i_1, \ldots, i_n)$ dos números $1, 2, \ldots, n$. Esta fórmula teórica só permite o cálculo *efectivo* do determinante se a dimensão da matriz for muito pequena. Por exemplo, se $n = 25$ o número de permutações possíveis é superior a 15 quatriliões (como é que se escreve este número?)! Se possuirmos uma máquina que calcule cada termo da expressão anterior num bilionésimo de segundo (coisa que nem remotamente os actuais computadores conseguem fazer), para calcular todas as parcelas necessitamos de 15 biliões de segundos, ou seja 400.000 anos!

Os problemas que a análise numérica pretende dar solução são geralmente originários das ciências naturais e sociais, da engenharia, das finanças, e, como foi dito, não podem, geralmente, ser resolvidos por processos analíticos.

Um dos primeiros e mais importantes modelos matemáticos para problemas da física foi estabelecido por Isaac Newton (1643-1729) para descrever o efeito da gravidade. De acordo com esse modelo, a força da gravidade exercida pela Terra num corpo de massa m tem a magnitude

$$F = G\frac{m \times m_t}{d^2},$$

onde m_t é a massa da Terra, d a distância entre os centros dos dois corpos e G a constante de gravitação universal. O modelo de Newton para a gravitação universal conduziu a ciência à formulação de muitos problemas cuja solução só pode ser obtida de forma aproximada, usualmente envolvendo a solução numérica de equações diferenciais.

Exemplo 1.1 (Problema dos três corpos) O problema dos três corpos consiste em determinar quais são os comportamentos possíveis de um sistema constituído por três corpos que interagem entre si através de uma força gravitacional newtoniana. Este problema não é difícil de pôr em equação e os espectaculares êxitos da mecânica clássica dos finais do século XIX sugeriam que a sua resolução, de interesse aparentemente académico, fosse uma questão de

tempo; o facto de não ser possível realizar os cálculos podia passar de mero detalhe técnico. Afinal de contas, o problema dos dois corpos (isto é, dois corpos que interagem por via da força gravitacional, como a Terra e o Sol) tinha uma solução muito simples, que era estudada no primeiro ano das universidades. O facto é que a solução analítica deste problema é impossível de obter! Resta-nos assim recorrer à solução numérica.

O estabelecimento das várias leis da física permitiu aos matemáticos e aos físicos obter modelos para a mecânica dos sólidos e dos fluidos. As engenharias mecânica e civil usam esses modelos como sendo a base para os mais modernos trabalhos em dinânica dos fluidos e em estruturas sólidas, e a análise numérica tornou-se uma ferramenta essencial para todos aqueles que pretendem efectuar investigação nessas áreas da engenharia. Por exemplo, a construção de estruturas modernas faz uso do chamado método dos elementos finitos para resolver as equações com derivadas parciais associadas ao modelo; a dinâmica dos fluidos computacional é actualmente uma ferramenta fundamental para, por exemplo, desenhar aviões; a elaboração de novos materiais é outro assunto que recorre, de forma intensa, a algoritmos numéricos. A análise numérica é pois uma área que tem assumido crescente importância no contexto das ciências da engenharia.

No processo de resolução de um problema físico podemos distinguir várias fases.

1. Formulação de um modelo matemático que descreve uma situação real. Tal formulação pode ser feita recorrendo a (sistemas de) equações algébricas, transcendentes, integrais, diferenciais, etc. É necessário ter muito cuidado nesta fase uma vez que a grande complexidade dos problemas físicos pode-nos obrigar a fazer simplificações no modelo, simplificações essas que não devem alterar grandemente o comportamento da solução.

2. Obtenção de um método numérico que permite construir uma solução aproximada para o problema. Um método numérico que possa ser usado para resolver o problema é traduzido por algoritmo que não é mais do que um completo e não ambíguo conjunto de passos que conduzem à solução do problema. Esta fase constitui o cerne da análise numérica. Dado um determinado método numérico, temos necessidade de saber em que condições as soluções por ele obtidas convergem para a solução exacta; em que medida pequenos erros de arredondadmento (e outros) poderão afectar a solução final; qual o grau de precisão da solução aproximada obtida, etc.

3. Programação automática do algoritmo. Nesta fase teremos necessidade de recorrer a uma linguagem de programação como o Fortran, o Pascal, o C++, entre outras. Mais recentemente é usual o recurso a programas como o Mathematica ou o Matlab.

Os algoritmos numéricos são quase tão antigos quanto a civilização humana. Os babilónios, vinte séculos antes de Cristo, já possuiam tabelas de quadrados de todos os inteiros entre 1 e 60. Os egípcios, que já usavam fracções, inventaram o chamado método da falsa posição para aproximar as raízes de uma equação. Esse método encontra-se descrito no papiro de Rhind, cerca de 1650 anos antes da era cristã.

Na Grécia antiga muitos foram os matemáticos que deram contributos para o impulso desta disciplina. Por exemplo, Arquimedes de Siracusa (278-212, a.C.) mostrou que

$$3\frac{10}{71} < \pi < 3\frac{1}{7}$$

e apresentou o chamado método da exaustão para calcular comprimentos, áreas e volumes de figuras geométricas. Este método, quando usado como método para calcular aproximações,

está muito próximo do que hoje se faz em análise numérica; por outro lado, foi também um importante precursor do desenvolvimento do cálculo integral por Newton e Gottfried Wilhelm von Leibniz (1646-1716).

Heron de Alexandria (∼10-∼75), no século I, deduziu um procedimento para determinar $\sqrt{a}$ da forma (como deduzir este método?)

$$x^{(n+1)} = \frac{1}{2}\left(x^{(n)} + \frac{a}{x^{(n)}}\right).$$

No ano 250, Diofanto de Alexandria (∼200-∼284) obteve um processo para a determinação das soluções de uma equação quadrática. Durante a Idade Média, os grandes contributos para o desenvolvimento da matemática algorítmica vieram, sobretudo, do médio oriente, Índia e China. O contributo maior foi, sem dúvida, a simplificação introduzida com a chamada numeração indo-árabe.

O aparecimento do cálculo e a criação dos logaritmos, no século XVII, vieram dar um grande impulso ao desenvolvimento de procedimentos numéricos. Os novos modelos matemáticos propostos não podiam ser resolvidos de forma explícita e assim tornava-se imperioso o desenvolvimento de métodos numéricos para obter soluções aproximadas. O próprio Newton criou vários métodos numéricos para a resolução de muitos problemas, métodos esses que possuem, hoje, o seu nome. Tal como Newton, muitos vultos da matemática dos séculos XVIII e XIX trabalharam na construção de métodos numéricos. De entre eles podemos destacar Leonhard Euler (1707-1783), Joseph-Louis Lagrange (1736-1813) e Johann Carl Friedrich Gauss (1777-1875).

Foi, no entanto, o aparecimento, na década de 40 do século XX, dos primeiros computadores que contribuiu decisivamente para o forte desenvolvimento da disciplina. Apesar de tanto Blaise Pascal (1623-1662) como Leibniz terem construído, já no séc. XVII, as primeiras máquinas de calcular e de Charles Babbage (1791-1871), milionário inglês, ter construído o que é considerado o primeiro computador (nunca funcionou!), foi apenas com o aparecimento do ENIAC, nos anos 40, que a ciência usufruiu, de facto, desses dispositivos de cálculo.

1.1 Erros absolutos e relativos

O processo de resolução numérica de um problema passa sempre pela substituição de um problema difícil para um problema mais simples que tenha a mesma solução ou, pelo menos, uma solução que se relaciona intimamente com a do problema original. As soluções numéricas são, por isso, aproximações dos resultados pretendidos. Os motivos para considerar essas aproximação podem ser de vária ordem e a precisão da solução final reflecte todo esse processo. Também há que ter em conta a incerteza nos dados iniciais, assim como as perturbações efectuadas ao longo dos cálculos, que podem ser amplificadas pelo algoritmo.

Exemplo 1.2 Pretende-se calcular a superície terrestre pela fórmula $A = 4\pi r^2$. Este processo involve uma aproximação. Primeiro, a Terra não é uma esfera. A esfera é uma idealização da sua verdadeira forma. Depois, o valor do raio da Terra é baseado em medições empíricas e cálculos anteriores. O valor de π requer a truncatura de um processo infinito. Finalmente, os valores dos dados iniciais e os resultados das operações aritméticas são arredondados durante os cálculos no computador.

O exemplo anterior mostra que a introdução de erros num determinado processo de cálculo pode ter várias causas. É nosso objectivo analisar quais são essas causas e estudar

mecanismos que nos permitam determinar limites superiores para os erros obtidos no final do processo de cálculo.

Definição 1.1 (Erro) *Seja $x \in \mathbb{R}^n$ um vector cujas componentes são desconhecidas e $\overline{x} \in \mathbb{R}^n$ um vector cujas componentes são aproximações para as componentes correspondentes de x. Chama-se* erro *de $\overline{x}$ (como aproximação a x), e representa-se por $e(\overline{x})$, à quantidade*

$$e(\overline{x}) = x - \overline{x}.$$

Na prática, o valor do erro é usado, geralmente, em norma. No caso real ($n = 1$), por exemplo, para a maioria dos problemas, não é relevante saber se o erro foi cometido por defeito ou por excesso. Vamos, então relembrar o conceito de norma vectorial.

Definição 1.2 (Norma) *Seja E um espaço vectorial (real ou complexo). A aplicação $\|\cdot\| : E \longrightarrow \mathbb{R}_0^+$ que verifica*

1. $\forall x \in E, \qquad \|x\| = 0 \Leftrightarrow x = 0,$
2. $\forall x \in E,\ \forall \lambda \in \mathbb{R}\ (ou\ \mathbb{C}), \qquad \|\lambda x\| = |\lambda|\|x\|,$
3. $\forall x, y \in E, \qquad \|x + y\| \leq \|x\| + \|y\|,$

é designada por norma.

Como consequência da propriedade 3 da definição anterior temos

$$\|u\| = \|u - v + v\| \leq \|u - v\| + \|v\|$$

e, portanto, $\|u\| - \|v\| \leq \|u - v\|$. Por outro lado, $-\|u\| + \|v\| \leq \|u - v\|$ e, como tal,

$$|\|u\| - \|v\|| \leq \|u - v\|.$$

Existem várias funções que verificam as três propriedades das normas vectoriais. Entre elas destacam-se as dadas no próximo exercício.

Exercício 1.1 Prove que as funções seguintes são normas em $\mathbb{R}^n$:

- $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, (norma um);
- $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$, (norma euclidiana);
- $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$, (norma do máximo).

A norma do máximo é também chamada norma de Pafnuty Lvovich Chebyshev (1821-1894).

Dado um vector $x \in \mathbb{R}^n$, prova-se facilmente que $\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty$. No entanto, também é possível provar que $\|x\|_1 \leq \sqrt{n}\|x\|_2$, $\|x\|_2 \leq \sqrt{n}\|x\|_\infty$, $\|x\|_1 \leq n\|x\|_\infty$. Isto significa que, fixado n, as diferentes normas diferem apenas numa constante multiplicativa e, como nesse sentido, dizem-se normas equivalentes: se uma delas se torna pequena, todas as outras serão proporcionalmente pequenas.

Vamos agora introduzir os conceitos de erro absoluto e relativo.

Definição 1.3 (Erro absoluto) *Seja $x \in \mathbb{R}^n$ um vector cujas componentes são desconhecidas e $\overline{x} \in \mathbb{R}^n$ um vector cujas componentes são aproximações para as componentes correspondentes de x. Chama-se* erro absoluto *de $\overline{x}$ à quantidade $\|e(\overline{x})\|$.*

Definição 1.4 (Erro relativo) *Seja $x \in \mathbb{R}^n$, $x \neq 0$, um vector cujas componentes são desconhecidas e $\overline{x} \in \mathbb{R}^n$ um vector cujas componentes são aproximações para as componentes correspondentes de x. Chama-se* erro relativo *de $\overline{x}$, e representa-se por $r(\overline{x})$, à quantidade*

$$r(\overline{x}) = \|e(\overline{x})\|/\|x\|.$$

Como na definição de erro relativo o valor de x não é conhecido, é usual considerar a aproximação $r(\overline{x}) \approx \|e(\overline{x})\|/\|\overline{x}\|$. Melhor ainda, atendendo a que

$$\|x\| \geq |\|\overline{x}\| - \|e(\overline{x})\||,$$

podemos considerar o majorante

$$r(\overline{x}) \leq \frac{\|e(\overline{x})\|}{|\|\overline{x}\| - \|e(\overline{x})\||}.$$

O erro relativo, atendendo a que é uma quantidade adimensionada, é muitas vezes representado sob a forma de percentagem. Note-se também que o erro relativo nos dá uma maior informação quanto à precisão da aproximação que o erro absoluto.

É com base nas definições de erro absoluto e erro relativo que iremos analisar os resultados numéricos que aparecerão como aproximações a valores que não conhecemos com exactidão. Suponhamos que pretendemos calcular o valor de uma função $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ para um dado argumento vectorial $x \in \mathbb{R}^n$ (valor exacto). O valor desejado é, pois, $F(x)$. No entanto, o que realmente é calculado é o valor de $\overline{F}(\overline{x})$, onde $\overline{x}$ é o vector das aproximações e $\overline{F}$ é a aproximação à função, o algoritmo. Assim, o erro total é dado por

$$F(x) - \overline{F}(\overline{x}) = \underbrace{F(\overline{x}) - \overline{F}(\overline{x})}_{\text{erro computacional}} + \underbrace{F(x) - F(\overline{x})}_{\text{erro de propagação}}.$$

O algoritmo não tem qualquer influência nos erros de propagação. Apenas influencia os erros computacionais, que são a soma dos erros de arredondamento com os erros de truncatura.

1.2 Erros de arredondamento e truncatura

Os dados de um determinado problema podem estar à partida afectados de imprecisões resultantes de medições incorrectas. Note-se que a escala de um instrumento de medição nos dá uma possibilidade de saber um limite superior para o erro com que vêm afectados os valores medidos. Por exemplo, com uma régua usual, a medição de uma distância de 2 mm pode vir afectada com um erro de 0,5 mm o que dá um erro relativo de 25%. Outra causa de erro resulta das simplificações impostas ao modelo matemático usado para descrever um determinado fenómeno físico. Por exemplo, é usual considerar que, para um dada problema, não há perdas de calor, o atrito é nulo, etc. Este tipo de erros fogem ao controlo do analista numérico e são muito difíceis de quantificar.

O aspecto que vamos agora analisar tem a ver com os erros que resultam da forma como representamos os números reais. O conjunto dos números reais $\mathbb{R}$ não pode ser representado numa máquina de precisão finita. Numa máquina só é possível representar um seu subconjunto finito $\mathbb{F}$. Os números desse conjunto $\mathbb{F}$ são chamados números de vírgula flutuante. Um número real x é geralmente truncado pela máquina dando origem a um novo número que (número de vírgula flutuante), que se designa por $\text{fl}(x)$. Em geral, $x \neq \text{fl}(x)$. Além disso, podemos ter $x_1 \neq x_2$ e $\text{fl}(x_i) = \text{fl}(x_2)$.

Usualmente, um computador guarda um número real na forma

$$x = (-1)^s \cdot (0{,}a_1a_2 \ldots a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t}, \qquad a_1 \neq 0, \tag{1.1}$$

onde s é 0 ou 1, conforme o sinal de x, β, inteiro positivo maior ou igual a 2, é a base adoptada pelo computador específico em que estamos a trabalhar, m é um inteiro chamado mantissa, cujo comprimento t é o número máximo de algarismos armazenados a_i, com $0 \leq a_i \leq \beta - 1$, e e um número inteiro chamado expoente. Os dígitos $a_1a_2 \ldots a_p$, com $p \leq t$, são chamados os p primeiros algarismos significativos de x.

O conjunto $\mathbb{F}$ fica completamente caracterizado à custa de 4 parâmetros: a base β, número de algarismos significativos t, e o intervalo de variação do expoente e, designado por $]L, U[$, com $L < 0$ e $U > 0$. Escrevemos então $\mathbb{F}(\beta, t, L, U)$. Em Matlab temos $\mathbb{F} = \mathbb{F}(2, 53, -1021, 1024)$. Note-se que, 53 algarismos significativos em base 2 correspondem a 15 algarismos significativos em base 10.

Para perceber melhor o que está em causa, consideremos, por exemplo, o número $x = 123{,}9346$. Este número não tem representação numa máquina de base decimal cuja mantissa só permita armazenar 6 dígitos. Temos assim necessidade de o aproximar por um outro que possa ser representado na referida máquina. Essa aproximação vai ser efectuada por um processo conhecido por arredondamento.

A forma de arredondar um número real é a usual. Como tal

$$x = 123{,}9346 \approx 123{,}935 = \overline{x},$$

e este novo valor já tem representação na máquina que estamos a usar sob a forma $0{,}123935 \times 10^2$.

Note-se que o arredondamento foi efectuado na terceira casa decimal e que

$$|e(\overline{x})| = |x - \overline{x}| = 0{,}0004 < 0{,}5 \times 10^{-3},$$

$$r(\overline{x}) = \frac{|e(\overline{x})|}{|x|} \approx 3{,}23 \times 10^{-6} < 5 \times 10^{-6}.$$

Se o arredondamento tivesse sido efectuado na segunda casa decimal vinha

$$x = 123{,}9346 \approx 123{,}93 = \overline{\overline{x}},$$

e assim

$$|e(\overline{\overline{x}})| = 0{,}0045 < 0{,}5 \times 10^{-2},$$

$$r(\overline{\overline{x}}) = \frac{|e(\overline{\overline{x}})|}{|x|} \approx 3{,}63 \times 10^{-5} < 5 \times 10^{-5}.$$

Daqui resultam as seguintes definições.

Definição 1.5 (Casa decimal correcta) *Seja $\overline{x} \in \mathbb{R}$ uma aproximação para $x \in \mathbb{R}$. Diz-se que $\overline{x}$ tem k* casas decimais correctas *se e só se $|e(\overline{x})| \leq 0{,}5 \times 10^{-k}$.*

Definição 1.6 (Algarismo significativo correcto) *Seja $\overline{x} \in \mathbb{R}$ uma aproximação para $x \in \mathbb{R}$. Diz-se que $\overline{x}$ tem k* algarismos significativos correctos *se e só se $r(\overline{x}) < 5 \times 10^{-k}$.*

Note-se que estas definições surgem por forma a que todo o número obtido a partir de um valor exacto por conveniente arredondamento tenha todas as suas casas decimais e todos os seus algarismos significativos correctos.

Consideremos, de novo, a máquina $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$. O erro que se comete na aproximação $x \approx \mathrm{fl}(x)$ é pequeno. Ele é dado por

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq 0{,}5\epsilon_M,$$

onde $\epsilon_M = \beta^{1-t}$ representa o zero da máquina e é definido como sendo o menor número que pode ser representado satisfazendo a

$$(1 + \epsilon_M) > 1.$$

Assim, uma máquina é tanto mais precisa quanto menor for o seu zero. Em Matlab, o valor de ϵ_M pode ser obtido com o comando `eps` e tem-se $\epsilon_M = 2^{-52} \approx 2.22 \times 10^{-16}$.

Notemos que, uma vez que $a_1 \neq 0$ em (1.1), o 0 não pertende a $\mathbb{F}$. Por outro lado, não é possível representar números arbitrariamente grandes ou arbitrariamente pequenos uma vez que L e U são finitos. O menor e o maior real positivo de $\mathbb{F}$ são dados, respectivamente, por

$$x_{min} = \beta^{L-1}, \qquad x_{max} = \beta^U(1 - \beta^{-t}).$$

Em Matlab estes valores podem ser obidos através dos comandos `realmin` e `realmax`. Um número positivo menor que x_{min} produz uma mensagem de underflow; um número positivo maior que x_{max} produz uma mensagem de overflow e armazena-se, em Matlab, na variável `Inf`.

Os elementos de $\mathbb{F}$ são mais densos próximos de x_{min} e menos densos quando se aproximam de x_{max}. O que se mantém constante é a distância relativa entre os números.

Finalmente, interessa observar que em $\mathbb{F}$ não existem formas indeterminadas como $0/0$ ou ∞/∞. Elas produzem o que se chama um not a number, denotado por `NaN` em Matlab.

Um problema sério a ter em conta em aritmética de vírgula flutuante é o chamado cancelamento subtractivo que ocorre, normalmente, quando se subtraem dois números com o mesmo sinal e magnitude semelhantes. Por exemplo,

$$1{,}92403 \times 10^2 - 1{,}92275 \times 10^2 = 1{,}28000 \times 10^1,$$

que é correcto, e representado de forma exacta, mas tem apenas três algarismos significativos.

Apesar da exactidão do resultado, o cancelamento implica, muitas vezes, perda de informação com a agravante de os algarismos perdidos no cancelamento serem *mais significativos* do que os perdidos no arredondamento. Por isso, é geralmente uma má ideia calcular uma quantidade pequena como a diferença de duas quantidades grandes. Somar séries alternadas, tais como

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

para $x < 0$, pode dar resultados catastróficos. Outro exemplo, pode ser dado no cálculo do desvio padrão. Como se sabe, a média e o desvio padrão de uma amostra $\{x_1, x_2, \ldots, x_n\}$ são dados por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{e} \quad \sigma = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}.$$

A fórmula equivalente para o desvio padrão

$$\sigma = \left(\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}\right)\right)^{\frac{1}{2}}$$

é muitas vezes usada mas o cancelamento subtractivo é mais perigoso que na fórmula anterior.

Os erros de truncatura ou de discretização são, por definição, os erros que surgem quando se passa de um processo infinito para um processo finito ou quando se substitui um processo contínuo por um discreto. A título de exemplo, considere-se o conhecido Teorema do Valor Médio, estabelecido por Joseph Louis Lagrange (1736-1813).

Teorema 1.1 (Valor Médio de Lagrange) *Se f for uma função contínua em $[a, b]$ e diferenciável em $]a, b[$ então existe pelo menos um $\xi \in]a, b[$ tal que*

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

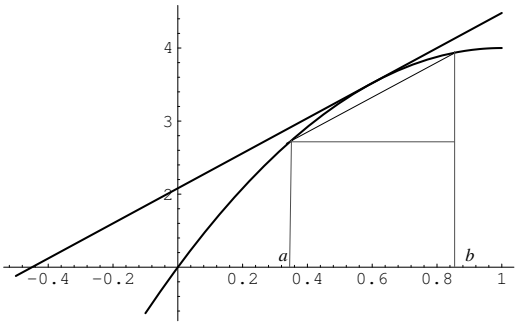

Figura 1.1: Teorema do Valor Médio.

Este resultado justifica o procedimento (muito comum) de substituir o cálculo da derivada de uma função definida num intervalo (pequeno) $[a, b]$ pela diferença dividida

$$f[a, b] = \frac{f(b) - f(a)}{b - a},$$

isto é, para um valor de h pequeno

$$f'(x) \approx \frac{f(x + h) - f(x)}{h}.$$

O erro cometido nesta aproximação é um erro de truncatura.

Também se comete um erro de truncatura quando se efectua a aproximação

$$e \approx \left(1 + \frac{1}{M}\right)^M.$$

Outro exemplo onde surgem este tipo de erros é dado pela chamada aproximação de Taylor que iremos considerar na próxima secção.

1.3 O polinómio de Taylor

Seja f uma função real definida num intervalo $[a, b] \subseteq \mathbb{R}$. Um problema que frequentemente se coloca é o de determinar uma função g definida em $[a, b]$ tal que $|f(x) - g(x)| < \epsilon$, para todo o $x \in [a, b]$, com $\epsilon > 0$ uma tolerância dada. A existência de solução para tal problema é dada pelo Teorema de Weierstrass, devido a Karl Wilhelm Theodor Weierstrass (1815-1897).

Teorema 1.2 (Weierstrass) *Seja f uma função contínua definida em $[a, b]$. Então para cada $\varepsilon > 0$ existe um polinómio p definido em $[a, b]$ tal que*

$$\max_{x \in [a,b]} |f(x) - p(x)| < \varepsilon.$$

Notemos a grande importância deste resultado. De acordo com ele, podemos ter a certeza que dada uma função contínua f qualquer existe sempre um polinómio p que está tão próximo de f quanto se queira. Assim sendo, este resultado legitima a aproximação polinomial, isto é, a tarefa de, dada uma função, procurar um polinómio que a aproxime. No entanto, o teorema anterior não nos diz como podemos construir esse polinómio; ele apenas garante a existência.

Consideremos agora o seguinte teorema, apresentado sem demonstração, devido a Brook Taylor (1685-1731).[1]

Teorema 1.3 (Taylor) *Se f admite derivadas contínuas até à ordem n (inclusivé) em $[a, b]$, isto é, se $f \in C^n([a, b])$, e se $f^{(n+1)}$ existir em $]a, b[$ então, para todo o $x, x_0 \in [a, b]$,*

$$f(x) = P_n(x; x_0) + R_n(x; x_0), \tag{1.2}$$

onde

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

e

$$R_n(x; x_0) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}, \qquad \xi \in I\{x, x_0\},$$

sendo $I\{x, x_0\}$ o intervalo aberto definido por x e x_0.

A (1.2) chamaremos fórmula de Taylor sendo $P_n(x; x_0)$ o polinómio de Taylor de f em torno do ponto x_0 e $R_n(x; x_0)$ o resto (de Lagrange) de ordem n (ou de grau $n + 1$). Se $x_0 = 0$ a (1.2) chamaremos fórmula de Maclaurin.[2]

[1] Taylor foi, entre outras coisas, o sucessor de Edmond Halley (1656-1742) como secretário da Royal Society. Publicou, em 1715, um livro intitulado *Methodus Incrementorum Directa & Inversa* no qual a sua expansão aparece descrita. O seu teorema foi enunciado em 1712.

[2] Colin Maclaurin (1698-1746) foi um menino prodígio sendo nomeado professor em Aberdeen com a idade de 19 anos. A sua expansão apareceu em 1742 no *Treatise on Fluxions*.

Atente-se ao grande interesse prático deste resultado que afirma que, mediante certas condições, uma função pode ser escrita como a soma de um polinómio com um resto. Escolhendo valores de x e x_0 tais que

$$\lim_{n\to+\infty} R_n(x;x_0) = 0, \tag{1.3}$$

temos que, a partir de um valor de n suficientemente grande, a função dada pode ser aproximada pelo seu polinómio de Taylor. Assim, qualquer operação a efectuar sobre a função (derivação, integração, etc.) poderá ser feita sobre o polinómio.

Notemos que a escolha dos valores de x e x_0 deverá ser feita de modo a que eles pertençam ao intervalo de convergência da série

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

designada por série de Taylor. Neste curso não iremos dar ênfase a esta questão.

O objectivo fundamental dos problemas que surgem neste contexto é o de determinar o menor valor de n que verifica

$$\max_{\xi\in I\{x,x_0\}} |R_n(x;x_0)| < \eta,$$

sendo $\eta > 0$ uma tolerância previamente fixada. Obtemos assim a aproximação

$$f(x) \approx P_n(x;x_0),$$

cujo erro não excede η. O valor de $R_n(x;x_0)$, sendo um erro absoluto uma vez que

$$|f(x) - P_n(x;x_0)| = |R_n(x;x_0)|,$$

é também designado erro de truncatura.

Exercício 1.2 Determine um valor aproximado de e^2 com 3 casas decimais correctas, usando a fórmula de Maclaurin aplicada à função $f(x) = e^x$.

Resolução: A função $f(x) = e^x$ é uma função analítica para todo o x real e atendendo a que $f^{(k)}(x) = e^x$ a série de Maclaurin de f é dada por

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Assim, fixando um valor de n, temos que

$$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!}$$

com

$$|R_n(x;0)| \leq \frac{e^x}{(n+1)!}|x^{n+1}| < \frac{3^x}{(n+1)!}|x^{n+1}|, \qquad x > 0.$$

Vamos então determinar qual o menor valor de n tal que

$$|R_n(2;0)| < \frac{3^2}{(n+1)!}|2^{n+1}| \leq 0{,}5 \times 10^{-3}.$$

Por tentativas...

$$n = 9 \quad \Rightarrow \quad \frac{3^2}{10!}2^{10} = 0{,}254 \times 10^{-2}$$

$$n = 10 \quad \Rightarrow \quad \frac{3^2}{11!}2^{11} = 0{,}462 \times 10^{-3}.$$

Logo a aproximação pedida é

$$e^2 \approx \sum_{k=0}^{10} \frac{x^k}{k!} = 7{,}38899470899 \approx 7{,}389.$$

1.4 Problemas

Exercício 1.3 Sejam x, y e z três quantidades exactas. Por arredondamento obtiveram-se as seguintes aproximações: $\overline{x} = 231$, $\overline{y} = 2{,}31$ e $\overline{z} = 23{,}147$.

1. Conte o número de casas decimais correctas nas aproximações e calcule limites superiores para o erro absoluto em cada uma delas. Compare os resultados e comente.

2. Conte o número de algarismos significativos correctos nas aproximações e calcule limites superiores para o erro relativo em cada uma delas. Compare os resultados e comente.

Exercício 1.4 (Matlab) A adição em vírgula flutuante não verifica a propriedade associativa. Para comprovar esse facto, calcule, usando o computador, a soma $x+y+z$ nas formas $x+(y+z)$ e $(x+y)+z$ para os casos em que:

1. $x = 1{,}0$; $y = -5{,}0$; $z = 6{,}0$;

2. $x = 1 \times 10^{20}$; $y = -1 \times 10^{20}$; $z = 1{,}0$.

Explique os resultados obtidos.

Exercício 1.5 (Matlab) Escreva um programa para obter os primeiros n termos da sucessão gerada pela equação de diferenças $x^{(k+1)} = 2{,}25x^{(k)} - 0{,}5x^{(k-1)}$, a partir dos valores iniciais $x^{(1)} = 1/3$ e $x^{(2)} = 1/12$.

1. Considere $n = 60$ e trace o gráfico semi-logarítmico dos resultados obtidos como função de k.

2. A solução exacta da equação de diferenças é dada por (prove)

   $$x^{(k)} = \frac{4^{1-k}}{3},$$

   que decresce, de forma monótona, quando k cresce. Explique porque é que o gráfico obtido na alínea anterior não está de acordo com este comportamento teórico.

Exercício 1.6 Estabeleça a fórmula de Taylor com o resto de Lagrange de ordem 3 para a função $f(x) = e^{-x}$, em torno do ponto $x_0 = 1$.

Exercício 1.7 (Matlab) Considere o desenvolvimento em série de Maclaurin

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Calcule e^{-12} usando este desenvolvimento. Repita novamente o cálculo mas usando o facto de $e^{-12} = \frac{1}{e^{12}}$. Compare os resultados e explique o sucedido.

Exercício 1.8 1. Calcule o polinómio de Taylor de grau 3 gerado pela função $f(x) = \cos x$ no ponto $x_0 = \frac{\pi}{4}$.

2. Usando o polinómio encontrado, calcule um valor aproximado de $\cos 47^o$.

3. Indique uma estimativa do erro absoluto cometido.

Exercício 1.9 Determine uma aproximação polinomial para $f(x) = \sinh(x)$ em $[0, 1]$, com erro inferior a 10^{-2}.

Exercício 1.10 Determine o número de termos que deve considerar em $\sum_{j=1}^{\infty}(-1)^{j+1}\frac{x^j}{j}$ para determinar uma aproximação para $f(x) = \log(1 + x),\ x \in [0, 1]$, com erro inferior a 10^{-3}.

Exercício 1.11 Calcule $\log(2)$ com erro inferior a 10^{-2}.

Exercício 1.12 O fluxo através de uma parte da camada fronteira num fluído viscoso é dado pelo integral definido

$$\int_0^{0,8} 1{,}4(1 - e^{-4x^2})dx.$$

Usando a fórmula de Taylor na função integranda, aproxime o valor do integral com quatro casas decimais correctas.

Exercício 1.13 Consideremos uma viga uniforme de comprimento L, suspensa, sujeita a uma carga uniformemente distribuída, W, e a uma força compressiva, P, em cada extremo. A deflexão, D, no ponto médio é dada por

$$D = \frac{WEI}{P^2}(\sec(0{,}5mL) - 1) - \frac{WL^2}{8P},$$

onde $m^2 = P/EI$, com E e I constantes. Usando o desenvolvimento em série de Maclaurin da função $y = \sec x$, prove que, quando a força gravítica tende a anular-se, a deflexão, D, tende para $\frac{5WL^4}{384EI}$.

Exercício 1.14 A lei dos gases perfeitos é dada por $PV = nrT$ e relaciona a pressão, P, o volume, V, a temperatura, T, e o número de moles, n, de um gás ideal. O número r nesta equação depende apenas do sistema de medição a usar. Suponhamos que foram efectuadas as seguintes experiências para testar a veracidade da lei usando o mesmo gás.

1. Consideraram-se $P = 1{,}0$ atmosferas, $n = 0{,}0042$ moles, $V = 0{,}10$ metros cúbicos e $r = 0{,}082$. Usando a lei, a temperatura do gás foi prevista como sendo

   $$T = \frac{PV}{nr} = \frac{1{,}0 \times 0{,}10}{0{,}082 \times 0{,}0042} = 290^o \text{ Kelvin } = 17^o \text{ Celsius.}$$

   Quando medimos a temperatura do gás verificámos ser 17^o Celsius.

2. A experiência anterior foi repetida usando os mesmos valores de r e n mas aumentando o pressão quatro vezes enquanto se reduziu o volume na mesma proporção. Como PV é constante, a temperatura prevista é de 17^o Celsius mas agora, ao medir a temperatura do gás, encontrámos o valor 32^o Celsius.

Será que a lei não é válida nesta situação?

Capítulo 2

Sistemas de equações lineares

O problema que pretendemos resolver neste capítulo consiste em determinar o valor de $x \in \mathbb{R}^n$ tal que $Ax = b$, sendo $b \in \mathbb{R}^n$ e $A \in \mathcal{M}_n(\mathbb{R})$, onde $\mathcal{M}_n(\mathbb{R})$ denota o conjunto das matrizes reais de ordem n. Para resolver este problema iremos supor que a matriz A é invertível ou, o que é equivalente, que o sistema é possível e determinado.

Há muitos sistemas físicos que podem ter como modelo sistemas de equações lineares. Suponhamos, por exemplo, um camião a atravessar uma ponte cuja estrutura é constituída por barras de ferro. O peso camião e da ponte são forças que são contrabalançadas pelas exercídas nas extremidades que seguram a ponte. Essas forças são propagadas ao longo de toda a estrutura e, em cada nodo (locais onde as barras de ferro seguram a estrutura) a resultante das forças deve ser nula. Se decompusermos as forças nas componentes horizontal (x) e vertical (y) temos, em cada nodo ($i = 1, 2, \ldots$), as equações:

$$\begin{array}{c} \text{soma } x \text{ das forças} = 0, \text{ no nodo } i, \\ \text{soma } y \text{ das forças} = 0, \text{ no nodo } i. \end{array}$$

As forças, em cada barra, podem assim ser determinadas. Como há forças conhecidas (peso do camião, peso das barras, etc), o sistema a resolver é não homogéneo.

A resolução de um problema envolvendo sistemas lineares pode dividir-se em três etapas:

1. formulação do modelo matemático (calcular a matriz A);
2. cálculo dos agentes exteriores (calcular o vector b);
3. resolução do sistema linear.

Os dois primeiros passos dependem, obviamente, do conhecimento do problema físico (tipo de material, leis físicas, etc); o terceiro passo pode ser equacionado e resolvido separadamente, usando um conveniente algoritmo matemático. Uma vez que este último passo aparece como pertencente a um algoritmo bastante mais vasto é essencial que seja calculado de forma eficiente.

Existem duas grandes classes de métodos para resolver sistemas de equações lineares: os métodos directos, que já foram estudados, em parte, na disciplina de álgebra linear e para os quais iremos fazer uma breve revisão; e os métodos iterativos que iremos estudar com mais pormenor, especialmente os métodos devidos a Gauss, Carl Gustav Jakob Jacobi (1804-1851) e Philipp Ludwig von Seidel (1821-1896). Antes porém, vamos apresentar algumas classes de matrizes que irão ser consideradas.

2.1 Classes de matrizes

Existem vários tipos de matrizes com relevância em aplicações práticas.

- Matrizes densas e matrizes esparsas

  Um matriz com muitos elementos nulos diz-se esparsa; caso contrário diz-se que a matriz é densa. Sistemas com matrizes esparsas modelam sobretudo problemas onde existem princípios de influência local. Note-se que, no caso da ponte, as equações em cada nodo apenas envolvem as barras que aí se encontram. O seu número é o mesmo quer a ponte tenha 50 metros e, digamos, 10 barras, ou 5 km e 1000 barras. Assim, para uma ponte grande, a maioria dos coeficientes da matriz são nulos.

  Como caso particular das matrizes esparsas temos as matrizes banda – uma matriz $A = (a_{ij})$ é uma matriz banda se $a_{ij} = 0$ para todo $|i-j| > \beta$, onde β é o comprimento de banda de A – e, dentro dessa classe, as chamadas matrizes tridiagonais, isto é, as matrizes com comprimento de banda $\beta = 3$. Por outras palavras, as matrizes banda são aquelas cujos elementos não nulos se concentram apenas num conjunto de diagonais paralelas à diagonal principal.

- Matrizes triangulares

  As matrizes triangulares são aquelas que possuem todos os seus elementos acima ou abaixo da diagonal principal iguais a zero. No primeiro caso, as matrizes dizem-se triangulares inferiores e no segundo triangulares superiores.

- Matrizes simétricas

  As matrizes simétricas são aquelas que coincidem com a sua transposta. Uma característica importante das matrizes simétricas é o facto de todos os seus valores próprios serem reais.

- Matrizes estritamente diagonal dominantes (EDD)

  Uma matriz $A = (a_{ij})_{i,j=1}^{n}$ diz-se estritamente diagonal dominante por linhas se

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

  e estritamente diagonal dominante por colunas se

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

  A matriz diz-se estritamente diagonal dominante se for estritamente diagonal dominante por linhas ou estritamente diagonal dominante por colunas.

- Matrizes simétricas e positivas definidas (SPD)

  Uma matriz A, diz-se simétrica e positiva definida se for simétrica e se, para todo o vector $x \in \mathbb{R}^n$, não nulo, se tem $x^T Ax > 0$. Nesse caso diz-se que $A \in$ SPD.

Exercício 2.1 Mostre que $A \in$ SPD se e só se todos os valores próprios de A são reais e positivos.

Resolução: Por definição, λ_i é valor próprio de $A \in \mathcal{M}_n(\mathbb{R})$ se e só se

$$Ax_i = \lambda_i x_i, \qquad x_i \in \mathbb{R}^n, \qquad x_i \neq 0,$$

sendo x_i o vector próprio de A associado a λ_i.

Suponhamos que $A \in$ SPD. Então $x^T Ax > 0$ é válida para todo o vector não nulo x e, em particular, para os vectores próprios x_i de A. Assim sendo,

$$0 < x_i^T Ax_i = x_i \lambda_i x_i = \lambda_i \|x\|_2^2 \Rightarrow \lambda_i > 0.$$

Suponhamos agora que todos os valores próprios de A são positivos. Então, pelo que foi visto, para todos os valores próprios x_i de A é válida a desigualdade $x_i^T Ax_i > 0$. Como A é uma matriz real simétrica, prova-se que existe uma base ortonormada de vectores próprios de A (teorema espectral). Nessa base, qualquer que seja o vector não nulo x, tem-se que existem constantes c_i, $i = 1, 2, ..., n$, onde n é a ordem da matriz A, tais que $x = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$. Então $Ax = c_1 Ax_1 + c_2 Ax_2 + \cdots + c_n Ax_n$ o que implica $Ax = c_1 \lambda_1 + c_2 \lambda_2 + \cdots + c_n \lambda_n$. Assim sendo, $x^T Ax = c_1^2 \lambda_1 + x_2^2 \lambda_2 + \cdots + c_n^2 \lambda_n > 0$, o que prova o pretendido.

2.2 Métodos directos: revisão

Consideremos, de novo, o problema de determinar o vector $x \in \mathbb{R}^n$ tal que $Ax = b$, sendo $b \in \mathbb{R}^n$ e $A \in \mathcal{M}_n(\mathbb{R})$. O primeiro tipo de métodos que iremos considerar resolver este problema são os chamados métodos directos. Estes métodos são aqueles que, supondo não haver erros de arredondamento ou quaisquer outros, nos permitem obter a solução exacta do problema num número finito de operações aritméticas.

Os métodos directos baseiam-se no processo de eliminação de Gauss que consiste em transformar o sistema $Ax = b$ num sistema equivalente $Ux = c$, onde U é uma matriz triangular superior, atravéz de operações elementares efectuadas na matriz ampliada (lembrar estes conceitos dados na disciplina de álgebra linear). O sistema a resolver pode ser escrito na forma

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n = c_1 \\ \qquad\quad u_{22}x_2 + \cdots + u_{2n}x_n = c_2 \\ \qquad\qquad\quad \ddots \qquad\quad \vdots \qquad\quad \vdots \\ \qquad\qquad\qquad\qquad u_{nn}x_n = c_n \end{cases}$$

e a sua resolução, caso $u_{ii} \neq 0$, $i = 1, \ldots, n$, é feita de acordo com o algoritmo seguinte.

Algoritmo 2.1 Resolução de um sistema triangular superior

Dados: c_i, $i = 1, \ldots, n$, e u_{ij}, $i = 1, \ldots, n$, $j = i, \ldots, n$

$x_n := c_n / u_{nn}$

Para i de $n-1$ até 1 fazer

$\quad x_i := \left(c_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) / u_{ii}$

Resultado: x_i, $i = 1, \ldots, n$

O método da eliminação de Gauss tem como desvantagem a alteração do valor dos termos independentes ($c \neq b$). Para contornar esse problema, temos o chamado método da triangularização. Este método consiste em decompor a matriz A do sistema a resolver na forma

$$A = LU,$$

em que L é uma matriz triangular inferior e U uma matriz triangular superior, com $u_{ii} = 1$, $i = 1, \ldots, n$. A forma de obter esta factorização foi igualmente vista na disciplina de álgebra linear e é chamada a factorização de Gauss. Após obtida a decomposição, a resolução do sistema é feita em duas etapas: resolver $Ly = b$; resolver $Ux = y$. Notemos que em cada etapa temos que resolver um sistema triangular.

Exercício 2.2 Supondo determinada a decomposição $A = LU$, obtenha o algoritmo que permita reslover o sistema $Ax = b$ pelo método da triangulação.

Em muitas situações práticas, o sistema linear a resolver é tridiagonal, isto é, a matriz A é da forma (suprimindo os zeros)

$$A = \begin{bmatrix} \beta_1 & \gamma_1 & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_{n-1} & \beta_{n-1} & \gamma_{n-1} \\ & & & \alpha_n & \beta_n \end{bmatrix},$$

ou, noutra notação,

$$A = \text{Tridiag}(\alpha, \beta, \gamma),$$

com $\alpha = [\alpha_2, \ldots, \alpha_n]^T$, $\beta = [\beta_1, \ldots, \beta_n]^T$ e $\gamma = [\gamma_1, \ldots, \gamma_{n-1}]^T$.

É fácil de demonstrar (ver Álgebra Linear) que a decomposição $A = LU$ é dada pelas matrizes

$$L = \begin{bmatrix} l_1 & & & & \\ \alpha_2 & l_2 & & & \\ & \ddots & \ddots & & \\ & & \alpha_{n-1} & l_{n-1} & \\ & & & \alpha_n & l_n \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} 1 & u_1 & & & \\ & 1 & u_2 & & \\ & & \ddots & \ddots & \\ & & & 1 & u_{n-1} \\ & & & & 1 \end{bmatrix},$$

com os elementos l_i, $i = 1, \ldots, n$ e u_i, $i = 1, \ldots, n-1$, dados de acordo com o seguinte algoritmo.

Algoritmo 2.2 Decomposição LU para matrizes tridiagonais

Dados: α_i, $i = 2, \ldots, n$, β_i, $i = 1, \ldots, n$, e γ_i, $i = 1, \ldots, n-1$
$l_1 := \beta_1$
Se $l_1 \neq 0$ então $u_1 := \gamma_1 / l_1$ caso contrário parar
Para i de 2 até $n-1$ fazer
    $l_i := \beta_i - \alpha_i u_{i-1}$
    Se $l_i \neq 0$ então $u_i := \gamma_i / l_i$ caso contrário parar
$l_n := \beta_n - \alpha_n u_{n-1}$
Resultado: l_i, $i = 1, \ldots, n$, e u_i, $i = 1, \ldots, n-1$

Como facilmente se pode demonstrar, a resolução do sistema linear $Ax = b$, com A uma matriz tridiagonal cuja decomposição $A = LU$ é dada pelo exercício anterior, pode ser efectuada de acordo com o seguinte algoritmo.

Algoritmo 2.3 Resolução de um sistema tridiagonal

Dados: matriz A e vector b
Determinar as matrizes L e U
($*$ Resolver $Ly = b$ $*$)
    $y_1 := b_1/l_1$
    Para i de 2 até n fazer
        $y_i := (b_i - \alpha_i y_{i-1})/l_i$
($*$ Resolver $Ux = y$ $*$)
    $x_n := y_n$
    Para i de $n-1$ até 1 fazer
        $x_i := y_i - u_i x_{i+1}$
Resultado: x_i, $i = 1, \ldots, n$

Prova-se o seguinte resultado.

Teorema 2.1 *Para uma matriz $A \in \mathcal{M}_n(\mathbb{R})$, a factorização $A = LU$ existe e é única se e só se as submatrizes principais de A forem não singulares.*

Quando, no processo da factorização de Gauss, encontramos um *pivot* nulo, podemos trocar de linhas por forma a evitar a divisão por zero. Uma possibilidade consiste em escolher, em cada passo da iteração, o *pivot* de módulo máximo. Obtém-se, assim, a factorização

$$PA = LU,$$

com P a matriz de permutação da linha efectuada.

Note-se que, se $A \in$ EDD ou $A \in$ SPD então a matriz A verifica as hipóteses do teorema anterior e, como tal, a factorização LU de A existe e é única. Mais ainda, se $A \in$ SPD pode obter-se a factorização de de A na forma $A = RR^T$, com R uma matriz triangular inferior com elementos diagonais positivos. Esta factorização é chamada factorização de Cholesky, em homenagem a André-Louis Cholesky (1875-1918). No caso 2×2, por exemplo,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 \\ r_{12} & r_{22} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}$$

implica que $r_{11} = \sqrt{a_{11}}$, $r_{12} = a_{12}/r_{11}$, $r_{22} = \sqrt{a_{22} - r_{12}^2}$.

2.3 Normas de matrizes. Condicionamento

Seja $\mathcal{M}_n(\mathbb{R})$ o conjunto das matrizes quadradas de ordem n com coeficientes reais. Como este conjunto é um espaço vectorial podemos nele definir uma norma.

Definição 2.1 (Norma matricial) *Seja $\|\cdot\|$ uma norma vectorial definida em $\mathbb{R}^n$. A aplicação $\|\cdot\| : \mathcal{M}_n(\mathbb{R}) \longrightarrow \mathbb{R}_0^+$ tal que, para todo o $A \in \mathcal{M}_n(\mathbb{R})$,*

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

é designada norma matricial subordinada à norma vectorial $\|\cdot\|$.

No entanto, a generalização para o caso rectangular Atendendo à definição anterior, podemos dizer que a norma de uma matriz mede o alongamento máximo que essa matriz provoca em qualquer vector não nulo em relação à correspondente norma vectorial.

Prova-se que a função estabelecida na definição anterior é uma norma, isto é, que verifica as seguintes propriedades:

1. $\forall A \in \mathcal{M}_n(\mathbb{R}), \qquad \|A\| = 0 \Leftrightarrow A = 0$ (matriz nula),
2. $\forall A \in \mathcal{M}_n(\mathbb{R}),\ \forall \lambda \in \mathbb{R}, \qquad \|\lambda A\| = |\lambda|\|A\|,$
3. $\forall A, B \in \mathcal{M}_n(\mathbb{R}), \qquad \|A + B\| \leq \|A\| + \|B\|,$

Prova-se também que as seguintes aplicações verificam as propriedades que caracterizam um norma, com $A \in \mathcal{M}_n(\mathbb{R})$ a matriz de elemento genérico a_{ij}:

- $\|A\|_1 = \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{n} |a_{ij}|;$
- $\|A\|_2 = \sqrt{\rho(A^T A)}$, com $\rho(A)$ o raio espectral de A, definido por

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

- $\|A\|_\infty = \max\limits_{i=1,\ldots,n} \sum\limits_{j=1}^{n} |a_{ij}|.$
- $\|A\|_F = \sqrt{\sum\limits_{i,j=1}^{n} a_{ij}^2}.$

A norma $\|\cdot\|_F$ é chamada norma de Frobenius, em homenagem a Ferdinand Georg Frobenius (1849–1917), e não é soburdinada a nenhuma norma vectorial.

Exercício 2.3 Considere a matriz

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & -2 & 1 \\ -2 & -1 & 0 \end{bmatrix}.$$

Calcule $\|A\|_1$, $\|A\|_\infty$ e $\|A\|_F$.

Resolução: Vê-se facilmente que

$$\|A\|_1 = \max\{|2| + |-1| + |-2|, |-1| + |-2| + |-1|, |0| + |1| + |0|\} = 5,$$

$$\|A\|_\infty = \max\{|2| + |-1| + |0|, |-1| + |-2| + |1|, |-2| + |-1| + |0|\} = 4$$

e

$$\|A\|_F = \sqrt{2^2 + (-1)^2 + 0^2 + (-1)^2 + (-2)^2 + 1^2 + (-2)^2 + (-1)^2 + 0^2} = 4.$$

Exercício 2.4 Mostre que se $A \in$ SPD então $\|A\|_2 = \lambda_{max}$, com λ_{max} o maior valor próprio da matriz A.

Resolução: Por definição

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

Como A é simétrica $A^T A = A^2$. Mas, $\rho(A^2) = \rho(A)^2$. De facto, se λ é valor próprio de $A \in \mathcal{M}_n(\mathbb{R})$,

$$Ax = \lambda x, \qquad x \in \mathbb{R}^n, \qquad x \neq 0.$$

Então

$$A^2 x = \lambda A x = \lambda^2 x, \qquad x \in \mathbb{R}^n, \qquad x \neq 0,$$

ou seja λ^2 é valor próprio de A^2. Temos, então, que

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

Como $A \in$ SPD, $\lambda_{max} > 0$ e portanto, $\lambda_{max} = \rho(A)$, o que prova o pretendido.

Considere-se, agora, o resultado seguinte.

Teorema 2.2 *Seja $\|\cdot\|$ uma norma em $\mathbb{R}^n$. Então, para $A \in \mathcal{M}_n(\mathbb{R})$, tem-se*

$$\|Ax\| \leq \|A\|\|x\|, \forall x \in \mathbb{R}^n,$$

em que $\|A\|$ é a norma de A subordinada à norma $\|\cdot\|$.

Demonstração: Notamos que se a norma $\|\cdot\|$ é subordinada a uma norma vectorial então, para $A \in \mathcal{M}_n(\mathbb{R})$, temos

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Ax\|}{\|x\|}, \qquad x \in \mathbb{R}^n, x \neq 0,$$

e portanto é válido o resultado. □

Teorema 2.3 *Para $A, B \in \mathcal{M}_n(\mathbb{R})$, tem-se $\|AB\| \leq \|A\|\|B\|$.*

Demonstração: Atendendo à desigualdade demonstrada no teorema anterior, temos

$$\|AB\| = \sup_{x \neq 0} \frac{\|ABx\|}{\|x\|} \leq \|A\| \sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} = \|A\|\|B\|. \quad \square$$

Consideremos um sistema possível e determinado $Ax = b$ e seja $\overline{b}$ o vector obtido a partir de b considerando perturbações numéricas nas suas componentes. Esta situação é frequente quando o vector dos termos independentes representa medições. Sejam $e(\overline{b})$ e $r(\overline{b})$ respectivamente os erros absoluto e relativo de $\overline{b}$. Vejamos de que modo este erros influenciam os erros absoluto e relativo de $\overline{x}$, sendo $\overline{x}$ a solução do sistema $A\overline{x} = \overline{b}$.

Este problema também poderia ser colocado de outra forma. Pretendemos resolver o sistema $Ax = b$ pelo método de eliminação de Gauss usando, por exemplo, a factorização $A = LU$. Atendendo aos possíveis erros de arredondamento, obtemos uma solução $\overline{x}$ tal

que $A\overline{x} = \overline{b} \neq b$. Queremos saber de que forma o erro (absoluto ou relativo) em $\overline{b}$ influencia o erro em $\overline{x}$.

Se o sistema $Ax = b$ é possível e determinado, então A é invertível e portanto $x = A^{-1}b$. Consideremos agora o sistema em que o vector dos termos independentes tem as componentes afectadas de erro, i.e, $A\overline{x} = \overline{b}$. Temos

$$\|r(\overline{x})\| = \frac{\|x - \overline{x}\|}{\|x\|} = \frac{\|A^{-1}(b - \overline{b})\|}{\|x\|}. \tag{2.1}$$

Como, pelo Teorema 2.2, $\|b\| = \|Ax\| \leq \|A\|\|x\|$, concluímos que

$$\|x\| \geq \|A\|^{-1}\|b\|.$$

Utilizando esta desigualdade em (2.1) deduzimos

$$\|r(\overline{x})\| \leq \frac{\|A^{-1}\|\|b - \overline{b}\|}{\|A\|^{-1}\|b\|} = \|A\|\|A^{-1}\|\|r(\overline{b})\|.$$

A

$$K(A) = \|A\|\|A^{-1}\|$$

chamamos número de condição[1] da matriz A. Então

$$\|r(\overline{x})\| \leq K(A)\|r(\overline{b})\|.$$

Note-se que se os dados do problema forem obtidos com a precisão da máquina, o erro relativo da solução x é dado por

$$\|r(\overline{x})\| \leq K(A)\epsilon_M.$$

Podemos então dizer que a solução calculada perde cerca de $\log_{10}(K(A))$ dígitos de precisão relativamente aos dados do problema.

Do exposto podemos afirmar que se o número de condição de A for pequeno pequenas perturbações no vector dos termos independentes conduzem a pequenas perturbações no vector solução. Neste caso dizemos que o sistema $Ax = b$ e a matriz A são bem condicionados. Se o número de condição for muito grande o sistema $Ax = b$ e a matriz A dizem-se mal condicionados.

Exercício 2.5 Determine o número de condição de $A = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix}$ relativamente às normas $\|\cdot\|_1$ e $\|\cdot\|_\infty$.

Resolução: Como $A^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 0{,}5 \end{bmatrix}$, temos:

- norma $\|\cdot\|_1$

  $\|A\|_1 = \max\{|1| + |2|, |0| + |2|\} = 3,\ \|A^{-1}\|_1 = \max\{|1| + |-1|, |0| + |0{,}5|\} = 2,$

  e, como tal, $K_1(A) = 6$;

[1] A ideia de número de condição de uma matriz foi introduzida, em 1948, por Alan Mathison Turing (1912-1954), o mesmo que fundou a ciência da computação teórica e que cujos trabalhos foram decisivos para o sucesso do projecto inglês que decifrou o código de encriptação da famosa máquina "Enigma", permitindo pôr fim à ocupação do Atlântico pelos submarinos alemães durante a II Guerra Mundial.

- norma $\|\cdot\|_\infty$

  $\|A\|_\infty = \max\{|1|+|0|, |2|+|2|\} = 4,\ \|A^{-1}\|_\infty = \max\{|1|+|0|, |-1|+|0{,}5|\} = 1{,}5,$

  e, como tal, $K_\infty(A) = 6$.

Apesar do número de condição ter dado o mesmo em ambos os casos, tal poderia não ter acontecido.

Uma matriz ter um número de condição muito elevado significa que é "quase singular". Por convenção, se A é singular, $K(A) = \infty$. É possível demonstrar que

$$\|A\|\|A^{-1}\| = \left(\sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}\right)\left(\inf_{x\neq 0} \frac{\|Ax\|}{\|x\|}\right)^{-1}$$

e, como tal, o número de condição mede o razão entre o alongamento máximo e o encolhimento máximo que uma matriz provoca em qualquer vector não nulo.

Quando se pretende resolver sistemas lineares mal condicionados usando métodos directos, a solução obtida vem, frequentemente, afectada de erro. Nessa medida, é usual considerar métodos mistos, isto é, métodos iterativos que consideram como aproximação inicial a solução obtida pelo método directo. O método iterativo funciona assim como um corrector do resultado obtido pelo método directo.

2.4 Métodos iterativos

Consideremos, de novo, o problema de determinar o valor de $x \in \mathbb{R}^n$ tal que $Ax = b$, sendo $b \in \mathbb{R}^n$ e $A \in \mathcal{M}_n(\mathbb{R})$ uma matriz invertível. Um método iterativo para resolver o sistema consiste em, partindo de uma aproximação inicial $x^{(0)}$ para a solução do sistema (que iremos denotar por x^*), gerar uma sucessão de vectores $\{x^{(k)}\}$ convergente para x^*.

Os métodos que iremos considerar pertencem à classe dos métodos do ponto fixo e são obtidos transformando o problema $Ax = b$ num outro, equivalente, da forma

$$x = Bx + g,$$

para uma determinada matriz de iteração B e um determinado vector g.

Para determinar B e g podemos, por exemplo, decompor a matriz A na forma

$$A = P - (P - A),$$

com P uma matriz invertível (mais simples que A), chamada pré-condicionador, e considerar

$$B = P^{-1}(P - A) \quad \text{e} \quad g = P^{-1}b.$$

De facto,

$$Ax = b \Leftrightarrow Px - (P - A)x = b \Leftrightarrow Px = (P - A)x + b \Leftrightarrow x = P^{-1}(P - A)x + P^{-1}b.$$

Com esta transformação podemos escrever o método iterativo na forma

$$\begin{array}{lcl} x^{(0)} & & \text{dado,} \\ x^{(k+1)} & = & Bx^{(k)} + g, \quad k = 0, 1, \ldots. \end{array} \tag{2.2}$$

Atendendo a que

$$x^{(k+1)} = Bx^{(k)} + g \Leftrightarrow Px^{(k+1)} = (P - A)x^{(k)} + b \Leftrightarrow P(x^{(k+1)} - x^{(k)}) = b - Ax^{(k)},$$

o método (2.2) pode igualmente ser definido por

$$\begin{array}{rcl} P\delta^{(k)} & = & -r^{(k)}, \\ \delta^{(k)} & = & x^{(k+1)} - x^{(k)}, \end{array}$$

com

$$r^{(k)} = Ax^{(k)} - b.$$

O processo iterativo termina quando se cumprirem os critérios de paragem estabelecidos. Os critérios mais comuns são:

1. Critério do erro absoluto: $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$;
2. Critério do erro relativo: $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon \|x^{(k)}\|$;
3. Critério do número máximo de iterações: $k = k_{max}$.

Os métodos iterativos são sobretudo usados para matrizes esparsas de grandes dimensões, que surgem frequentemente em problemas de análise de circuitos ou cálculos de estruturas. Para esses problemas, os métodos iterativos são competitivos face aos métodos directos. Para matrizes densas ou de pequena dimensão os métodos directos são mais vantajosos.

Exemplo 2.1 Consideremos o seguinte sistema

$$Ax = b \Leftrightarrow \left\{ \begin{array}{rcrcrcr} 10x_1 & - & 2x_2 & + & 5x_3 & = & 13 \\ -x_1 & + & 3x_2 & - & x_3 & = & 1 \\ x_1 & + & 4x_2 & + & 2x_3 & = & 7 \end{array} \right.$$

que tem solução única $[x_1, x_2, x_3]^T = [1, 1, 1]^T$. Para converter o sistema na forma equivalente $x = Bx + g$ façamos

$$x = Bx + g \Leftrightarrow \left\{ \begin{array}{rcl} x_1 & = & \frac{1}{10}\left[2x_2 - 5x_3 + 13\right] \\ x_2 & = & \frac{1}{3}\left[x_1 + x_3 + 1\right] \\ x_3 & = & \frac{1}{2}\left[-x_1 - 4x_2 + 7\right] \end{array} \right. .$$

Neste caso temos que a matriz B de iteração e o vector g são dados por

$$B = \begin{bmatrix} 0 & 2/10 & -5/10 \\ 1/3 & 0 & 1/3 \\ -1/2 & -4/2 & 0 \end{bmatrix} \quad \text{e} \quad g = \begin{bmatrix} 13/10 \\ 1/3 \\ 7/2 \end{bmatrix}.$$

O método iterativo é dado na forma

$$x^{(k+1)} = Bx^{(k)} + g \Leftrightarrow \left\{ \begin{array}{rcl} x_1^{(k+1)} & = & \frac{1}{10}\left[2x_2^{(k)} - 5x_3^{(k)} + 13\right] \\ x_2^{(k+1)} & = & \frac{1}{3}\left[x_1^{(k)} + x_3^{(k)} + 1\right] \\ x_3^{(k+1)} & = & \frac{1}{2}\left[-x_1^{(k)} - 4x_2^{(k)} + 7\right] \end{array} \right. , \qquad k = 0, 1, \ldots,$$

sendo $x^{(0)} = [x_1^{(0)}, x_2^{(0)}, x_3^{(0)}]^T$ um valor dado.

O método descrito no exemplo anterior é conhecido por método de Jacobi e é um dos métodos que iremos estudar. Este método, quando aplicado ao sistema $Ax = b$, com $a_{ii} \neq 0$, $i = 1, \ldots, n$, pode ser dado por

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[- \sum_{j=1, j \neq i}^{n} a_{ij} x_j^{(k)} + b_i \right], \qquad i = 1, \ldots, n. \tag{2.3}$$

Exercício 2.6 Escreva um algoritmo para determinar a solução aproximada de $Ax = b$ pelo método de Jacobi.

Um melhoramento ao método de Jacobi pode ser dado de acordo com o próximo exemplo.

Exemplo 2.2 Consideremos, de novo, o sistema linear dado no exemplo anterior e o processo iterativo

$$\begin{cases} x_1^{(k+1)} &= \frac{1}{10} \left[2x_2^{(k)} - 5x_3^{(k)} + 13 \right] \\ x_2^{(k+1)} &= \frac{1}{3} \left[x_1^{(k+1)} + x_3^{(k)} + 1 \right] \\ x_3^{(k+1)} &= \frac{1}{2} \left[-x_1^{(k+1)} - 4x_2^{(k+1)} + 7 \right] \end{cases}, \qquad k = 0, 1, \ldots,$$

sendo $x^{(0)} = [x_1^{(0)}, x_2^{(0)}, x_3^{(0)}]^T$ um valor dado. Como se pode ver, este método usa as componentes da nova aproximação logo após estas terem sido calculadas. Neste caso a matriz de iteração do método já não é tão simples de escrever.

O método descrito no exemplo anterior é chamado método de Gauss-Seidel. Quando aplicado à resolução numérica de $Ax = b$, com $a_{ii} \neq 0$, $i = 1, \ldots, n$, o método é dado por

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[- \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \right], \qquad i = 1, \ldots, n. \tag{2.4}$$

Exercício 2.7 Escreva um algoritmo para determinar a solução aproximada de $Ax = b$ pelo método de Gauss-Seidel.

2.5 Convergência dos métodos iterativos

Vamos agora abordar a questão da convergência do método iterativo (2.2).

Definição 2.2 (Convergência) *Seja x^* a solução (única) de $Ax = b$ e $\{x^{(k)}\}$ uma sucessão de aproximações obtida pelo método (2.2). O método diz-se* convergente *se*

$$\lim_{k \to +\infty} \|e^{(k)}\| = \lim_{k \to +\infty} \left\| x^* - x^{(k)} \right\| = 0.$$

Caso contrário o método diz-se divergente. *A $e^{(k)}$ chama-se* erro (absoluto) *da iteração k.*

Notemos o seguinte resultado cuja demonstração é muito simples.

Teorema 2.4 *O método iterativo (2.2) converge, qualquer que seja a aproximação inicial $x^{(0)}$, se e só se*

$$\lim_{k \to +\infty} \|B^k\| = 0.$$

Demonstração: De facto, atendendo a que $x^* = Bx^* + g$, tem-se que

$$e^{(k)} = x^* - x^{(k)} = Bx^* - Bx^{(k-1)} = Be^{(k-1)}, \qquad k = 1, 2, \ldots.$$

Assim

$$e^{(k)} = B^k e^{(0)}, \qquad k = 1, 2, \ldots, \tag{2.5}$$

e, como tal,

$$\lim_{k\to+\infty} \|e^{(k)}\| = 0 \Leftrightarrow \lim_{k\to+\infty} \|B^k\| = 0,$$

o que prova o pretendido. □

Outro aspecto importante a determinar quando se lida com métodos iterativos tem a ver com a determinação de majorantes para o erro cometido. O teorema seguinte é, nesse sentido, muito importante. Além disso, estabelece uma condição suficiente de convergência mais útil que a referida no teorema anterior.

Teorema 2.5 *Se $\|B\| < 1$ então o método iterativo (2.2) converge, qualquer que seja a aproximação inicial escolhida, e tem-se que*

$$\|e^{(k)}\| \leq \|B\|^k \|e^{(0)}\|, \qquad k = 1, 2, \ldots. \tag{2.6}$$

Demonstração: Considerando normas em (2.5) obtemos (2.6). Tomando limites e atendendo a que $\|B\| < 1$ concluímos que o método é convergente. □

Notemos que, se $B \in$ SPD, a expressão (2.6) pode ser substituída por

$$\|e^{(k)}\| \leq \rho(B)^k \|e^{(0)}\|, \qquad k = 1, 2, \ldots. \tag{2.7}$$

Por outro lado, se for conhecido um valor aproximado de $\|B\|$ (ou $\rho(B)$), de (2.6) (ou (2.7)) poder-se-á deduzir o número mínimo de iterações k_{min} necessárias para reduzir o erro inicial $\|e^{(0)}\|$ de um factor ε. Com efeito, k_{min} será o menor inteiro positivo para o qual $\|B\|^{k_{min}} \leq \varepsilon$ (ou $\rho(B)^{k_{min}} \leq \varepsilon$).

O resultado anterior dá-nos apenas uma condição suficiente de convergência. Para estabelecer uma condição necessária e suficiente de convergência temos que usar a noção de raio espectral de uma matriz.

Pelo Exercício 2.25 podemos concluir que $\rho(A) \leq \|A\|$, o que permite estabelecer o seguinte teorema.

Teorema 2.6 *O método iterativo (2.2) converge, qualquer que seja a aproximação inicial $x^{(0)}$ escolhida, se e só se $\rho(B) < 1$.*

Demonstração: Se $\rho(B) < 1$ pode demonstrar-se que existe uma norma matricial tal que $\|B\| < 1$. Logo, pelo teorema anterior, o método converge.

Falta provar que se o método convergir então o raio espectral da matriz de iteração é menor que um. Vamos demonstrar este facto provando que se o raio espectral de B for maior ou igual a um podemos definir um processo iterativo de forma (2.2) divergente. De facto, se $\rho(B) \geq 1$ existe um valor próprio λ de B tal que $|\lambda| \geq 1$. Seja z o vector próprio associado a esse valor próprio. Considerando, em (2.2), a aproximação inicial $x^{(0)} = z$ e $g = z$ temos

$$x^{(1)} = Bz + z = \lambda z + z = (1 + \lambda)z.$$

Repetindo o processo temos, sucessivamente,

$$x^{(2)} = Bx^{(1)} + z = (1 + \lambda + \lambda^2)z, \quad \ldots, \quad x^{(k)} = Bx^{(k-1)} + z = \left(\sum_{j=0}^{k} \lambda^j \right) z.$$

Como $|\lambda| \geq 1$ concluímos que o método iterativo assim definido é divergente. □

Vamos agora estudar a convergência dos métodos de Jacobi e Gauss-Seidel. Seja $Ax = b$, com $A \in \mathcal{M}_n(\mathbb{R})$ uma matriz invertível. Considerando

$$A = P - (P - A), \qquad \text{com } P \text{ uma matriz invertível,} \tag{2.8}$$

podemos definir (como vimos) o método iterativo

$$\begin{array}{lcl} x^{(0)} & & \text{dado} \\ x^{(k+1)} & = & Bx^{(k)} + g, \quad k = 0, 1, \ldots, \end{array} \tag{2.9}$$

com $B = P^{-1}(P - A)$ e $g = P^{-1}b$.

A escolha dos diferentes métodos iterativos depende da forma como se define a partição (2.8), isto é, como se define o pré-condicionador P. Para definir os métodos de Jacobi e Gauss-Seidel, considera-se $A = D - L - U$, em que

$$D = \begin{bmatrix} a_{ii} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix},$$

$$L = \begin{bmatrix} 0 & & & \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}, \qquad U = \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_{n-1,n} \\ & & & 0 \end{bmatrix}.$$

Método de Jacobi. Para definir o método de Jacobi, temos sucessivamente

$$Ax = b \Leftrightarrow (D - L - U)x = b \Leftrightarrow Dx = (L + U)x + b.$$

Caso D seja invertível, temos que

$$x = D^{-1}(L + U)x + D^{-1}b.$$

O método de Jacobi é assim dado por (2.9), com $B = D^{-1}(L + U)$ e $g = D^{-1}b$, ou seja, é o método iterativo que resulta da escolha de $P = D$ como pré-condicionador.

Método de Gauss-Seidel. Para definir o método de Gauss-Seidel, consideramos

$$Ax = b \Leftrightarrow (D - L - U)x = b \Leftrightarrow (D - L)x = Ux + b.$$

Caso $D - L$ seja invertível ou, o que é equivalente, caso D seja invertível, temos que

$$x = (D - L)^{-1}Ux + (D - L)^{-1}b.$$

O método de Gauss-Seidel é assim dado por (2.9), com $B = (D-L)^{-1}U$ e $g = (D-L)^{-1}b$, ou seja, é o método iterativo que resulta da escolha de $P = D - L$ como pré-condicionador.

O estudo da convergência destes métodos iterativos pode ser efectuado de acordo com os resultados estabelecidos na secção anterior. Assim, se D for invertível, temos que, qualquer que seja a aproximação inicial escolhida:

- Método de Jacobi
  1. $\rho(D^{-1}(L+U)) < 1 \Leftrightarrow$ o método converge;
  2. $\|D^{-1}(L+U)\| < 1 \Rightarrow$ o método converge;
- Método de Gauss-Seidel
  1. $\rho((D-L)^{-1}U) < 1 \Leftrightarrow$ o método converge;
  2. $\|(D-L)^{-1}U\| < 1 \Rightarrow$ o método converge.

Para o caso particular dos métodos de Jacobi e Gauss-Seidel existe um resultado de convergência específico que pode ser útil na prática. Esse resultado é dado no próximo teorema, cuja demonstração irá ser feita apenas para o caso do método de Jacobi.

Teorema 2.7 *Se A é uma matriz estritamente diagonal dominante por linhas, então os métodos de Jacobi e Gauss-Seidel convergem para a única solução do sistema $Ax = b$, qualquer que seja a aproximação inicial escolhida.*

Demonstração: Vamos efectuar a demonstração apenas para o caso do método de Jacobi. A matriz B de iteração do método é dada por

$$B = D^{-1}(L+U) = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{bmatrix}.$$

Como A é estritamente diagonal dominante por linhas temos que

$$\|B\|_\infty = \max_{i=1,\ldots,n} \left\{ \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\} < 1.$$

Assim, pelo Teorema 2.5, temos que o método de Jacobi converge para a solução de $Ax = b$, qualquer que seja a aproximação inicial escolhida. □

Note-se que o teorema anterior nos dá apenas uma condição suficiente de convergência. Assim, se a matriz do sistema a for estritamente diagonal dominante por linhas os métodos de Jacobi e Gauss-Seidel irão gerar sucessões de aproximações convergentes para a sua solução; caso contrário, nada poderemos afirmar quanto à convergência dessas sucessões.

Exercício 2.8 Mostre que se A é uma matriz estritamente diagonal dominante por colunas, então o método de Jacobi converge para a única solução do sistema $Ax = b$, qualquer que seja a aproximação inicial escolhida.

2.6 O método dos mínimos quadrados

Desde a sua primeira aplicação a um problema de astronomia por Gauss[2], o método dos mínimos quadrados tem vindo a ser aplicado num vasto conjunto de situações tanto no campo da ciência como no da engenharia.

[2] A questão de saber a quem deve ser dado o crédito do método dos mínimos quadrados deu azo a uma famosa disputa entre Gauss, que o inventou, por volta de 1790, e Legendre, que o publicou primeiro, em 1805 (no mesmo ano em que Gauss inventou a transformada rápida de Fourier e que também não publicou).

Consideremos o sistema linear $Ax = b$ onde o número de equações excede o número de incógnitas, isto é, um sistema linear onde A é uma matriz do tipo $m \times n$, com $m > n$. Generalizando a notação usada durante o capítulo diremos que $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, com $m > n$. Os sistemas deste tipo são, em geral (mas nem sempre) impossíveis. No entanto, eles surgem em muitas aplicações práticas e, por isso, vão merecer a nossa atenção.

Regressão linear. Suponhamos duas grandezas físicas x e y relacionadas pela expressão $y = a + bx$, em que a e b são parâmetros a determinar. Suponhamos que foram efectuados seis pares de medições (x_i, y_i), $i = 1, 2, \ldots, 6$. Ficamos assim com o sistema (geralmente impossível)

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}.$$

Determinação de cotas em topografia. Obtiveram-se as cotas x_i de um conjunto de pontos $i = 1, 2, 3, 4$ relativamente uns aos outros, tendo-se tomado a cota do ponto $i = 4$ como referência. As relações obtidas foras as seguintes:

$$\begin{cases} x_1 - x_2 = 1{,}0 \\ x_2 - x_3 = 2{,}0 \\ x_3 - x_4 = 1{,}5 \\ x_1 - x_4 = 3{,}0 \\ x_1 - x_3 = 2{,}8 \\ x_2 - x_4 = 3{,}0 \end{cases}.$$

Em notação matricial, e fazendo $x_4 = 0$, temos o sistema impossível

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1{,}0 \\ 2{,}0 \\ 1{,}5 \\ 3{,}0 \\ 2{,}8 \\ 3{,}0 \end{bmatrix}. \tag{2.10}$$

Ajustar uma curva a um conjunto de pontos dados. São dadas as coordenadas (x_i, y_i), $i = 1, 2, \ldots, m$, de m pontos e pretende-se ajustar uma curva do tipo

$$f(x) = c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_n\phi_n(x),$$

em que $\phi_1, \phi_2, \ldots, \phi_n$ são funções conhecidas, sendo os c_j, com $j = 1, 2, \ldots, n$, parâmetros a determinar. Obtemos assim o sistema

$$Ax = b \Leftrightarrow \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_m) & \phi_2(x_m) & \ldots & \phi_n(x_m) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Como, em geral, $m \gg n$, o sistema é, normalmente, impossível.

Como vimos temos muitas vezes que considerar sistemas $Ax = b$, com $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, $m > n$, impossíveis. Nesses casos o resíduo $Ax - b$ é diferente de zero, isto é,

$$r(x) = Ax - b \neq 0, \qquad \forall x \in \mathbb{R}^n.$$

Como não existe $x \in \mathbb{R}^n$ que torna $r(x) = 0$, vamos determinar o vector $\widehat{x} \in \mathbb{R}^n$ que minimiza a norma do resíduo, isto é, que minimiza $\|r(x)\|$. O vector $\widehat{x}$ nestas condições é chamado a solução de $Ax = b$ (no sentido) dos mínimos quadrados, pois, normalmente, calcula-se o vector que minimiza $\|r(x)\|^2$.

O seguinte teorema, cuja demonstração será omitida, estabelece a existência e unicidade de solução do problema dos mínimos quadrados.

Teorema 2.8 *Para $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ e $b \in \mathbb{R}^m$ existe uma única solução de $Ax = b$ no sentido dos mínimos quadrados se e só se* $\text{car}(A) = n$.

A determinação da solução dos mínimos quadrados, caso exista, é feita de acordo com o seguinte teorema.

Teorema 2.9 *Para $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ e $b \in \mathbb{R}^m$ temos que $\widehat{x} \in \mathbb{R}^n$ é a solução de $Ax = b$ no sentido dos mínimos quadrados se e só se for a solução de $A^T A\widehat{x} = A^T b$.*

Exercício 2.9 Determine a solução dos mínimos quadrados do sistema $Ax = b$ dado por (2.10).

Resolução: Para este problema temos que

$$A^T A = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

e

$$A^T b = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1{,}0 \\ 2{,}0 \\ 1{,}5 \\ 3{,}0 \\ 2{,}8 \\ 3{,}0 \end{bmatrix} = \begin{bmatrix} 6{,}8 \\ 4{,}0 \\ -3{,}3 \end{bmatrix}.$$

Assim, o vector $\widehat{x} = \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{bmatrix}$, solução dos mínimos quadrados do sistema, é dado por

$$A^T A\widehat{x} = A^T b \Leftrightarrow \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{bmatrix} = \begin{bmatrix} 6{,}8 \\ 4{,}0 \\ -3{,}3 \end{bmatrix}.$$

Para resolver este sistema, vamos usar o método de eliminação de Gauss. Temos sucessivamente

$$\left[\begin{array}{ccc|c} \boxed{3} & -1 & -1 & 6{,}8 \\ -1 & 3 & -1 & 4{,}0 \\ -1 & -1 & 3 & -3{,}3 \end{array}\right] \xrightarrow[L_3 = 3L_3 + L_1]{L_2 = 3L_2 + L_1}$$

$$\left[\begin{array}{ccc|c} 3 & -1 & -1 & 6{,}8 \\ 0 & \boxed{8} & -4 & 18{,}8 \\ 0 & -4 & 8 & -3{,}1 \end{array}\right] \xrightarrow[L_3 = L_3 + (1/2)L_2]{} \left[\begin{array}{ccc|c} 3 & -1 & -1 & 6{,}8 \\ 0 & 8 & -4 & 18{,}8 \\ 0 & 0 & \boxed{6} & 6{,}3 \end{array}\right].$$

Passando à fase ascendente conclui-se imediatamente que

$$\widehat{x} = \left[\begin{array}{c} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{array}\right] = \left[\begin{array}{c} 3{,}575 \\ 2{,}875 \\ 1{,}05 \end{array}\right].$$

Consideremos os dados $\{(x_i, y_i), i = 0, \ldots, m\}$, que pretendemos ajustar a um modelo definido à custa de um número de parâmetros muito inferior ao número de dados. Uma situação muito usual consiste em considerar o modelo como sendo um polinómio

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

com $a_0, \ldots, a_n$ os parâmetros a determinar. Temos então que resolver o sistema

$$\left\{\begin{array}{lcl} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n & = & y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n & = & y_1 \\ & \vdots & \\ a_0 + a_1 x_m + a_2 x_m^2 + \cdots + a_n x_m^n & = & y_m \end{array}\right. .$$

que pode ser escrito na forma matricial

$$Ax = b \Leftrightarrow \left[\begin{array}{cccc} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & \cdots & x_m^n \end{array}\right] \left[\begin{array}{c} a_0 \\ a_1 \\ \vdots \\ a_n \end{array}\right] = \left[\begin{array}{c} y_0 \\ y_1 \\ \vdots \\ y_m \end{array}\right].$$

Com $n \ll m$, este sistema é, em geral, impossível. Como tal, $Ax - b \neq 0$. O problema dos mínimos quadrados, como vimos, consiste na determinação dos valores de $\widehat{x} = [\widehat{a}_0, \ldots, \widehat{a}_n]^T$ que minimizam a norma do quadrado dos resíduos, isto é, $\|Ax - b\|^2$. De acordo com o Teorema 2.9, o vector $\widehat{x}$ que torna mínima esta norma é a solução de $A^T A\widehat{x} = A^T b$, que, para este exemplo, pode ser escrito na forma

$$\left[\begin{array}{cccc} m+1 & \displaystyle\sum_{i=0}^{m} x_i & \cdots & \displaystyle\sum_{i=0}^{m} x_i^n \\ \displaystyle\sum_{i=0}^{m} x_i & \displaystyle\sum_{i=0}^{m} x_i^2 & \cdots & \displaystyle\sum_{i=0}^{m} x_i^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \displaystyle\sum_{i=0}^{m} x_i^n & \displaystyle\sum_{i=0}^{m} x_i^{n+1} & \cdots & \displaystyle\sum_{i=0}^{m} x_i^{2n} \end{array}\right] \left[\begin{array}{c} \widehat{a}_0 \\ \widehat{a}_1 \\ \vdots \\ \widehat{a}_n \end{array}\right] = \left[\begin{array}{c} \displaystyle\sum_{i=0}^{m} y_i \\ \displaystyle\sum_{i=0}^{m} y_i x_i \\ \vdots \\ \displaystyle\sum_{i=0}^{m} y_i x_i^n \end{array}\right]. \tag{2.11}$$

Estas equações são chamadas equações normais. Resolvendo as equações normais, obtemos os valores de $\widehat{a}_0, \ldots, \widehat{a}_n$ e, como tal, o polinómio dos mínimos quadrados.

Notemos que, no sistema de equações normais, a matriz é simétrica. Além disso, também se pode mostrar que, caso os pontos x_i, $i = 0, \ldots, m$, sejam distintos, é não singular. Assim sendo, o problema da determinação do polinómio dos mínimos quadrados tem solução única.

O problema da determinação do polinómio dos mínimos quadrados poderia ser colocado da seguinte forma alternativa: determinar os parâmetros $\widehat{a}_0, \ldots, \widehat{a}_n$ por forma a que

$$\phi(\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_n) = \min_{a_i, i=0,\ldots,n} \Phi(a_0, a_1 \ldots, a_n),$$

com

$$\Phi(a_0, a_1 \ldots, a_n) = \sum_{i=0}^{m} (y_i - a_0 - a_1 x_i \cdots - a_n x_i^n)^2.$$

Consideremos as observações $\{(x_i, y_i), i = 0, \ldots, m\}$ que correspondem a pontos do plano que pretendemos ajustar a uma recta da forma $p_1(x) = \widehat{a}_0 + \widehat{a}_1 x$. A questão que se coloca é a de determinar os valores $\widehat{a}_0$ e $\widehat{a}_1$ por forma a

$$\Phi(\widehat{a}_0, \widehat{a}_1) = \min_{a_0, a_1} \sum_{i=0}^{m} (y_i - a_0 - a_1 x_i)^2.$$

O ponto $(\widehat{a}_0, \widehat{a}_1)$ onde esta função atinge o mínimo satisfaz as condições

$$\left\{ \begin{array}{l} \dfrac{\partial \Phi}{\partial a_0}(\widehat{a}_0, \widehat{a}_1) = 0 \\ \dfrac{\partial \Phi}{\partial a_1}(\widehat{a}_0, \widehat{a}_1) = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \displaystyle\sum_{i=0}^{m} (y_i - \widehat{a}_0 - \widehat{a}_1 x_i) = 0 \\ \displaystyle\sum_{i=0}^{m} (y_i - \widehat{a}_0 - \widehat{a}_1 x_i) x_i = 0 \end{array} \right.$$

Temos então um sistema linear para resolver da forma

$$\left[\begin{array}{cc} m+1 & \displaystyle\sum_{i=0}^{m} x_i \\ \displaystyle\sum_{i=0}^{m} x_i & \displaystyle\sum_{i=0}^{m} x_i^2 \end{array} \right] \left[\begin{array}{c} \widehat{a}_0 \\ \widehat{a}_1 \end{array} \right] = \left[\begin{array}{c} \displaystyle\sum_{i=0}^{m} y_i \\ \displaystyle\sum_{i=0}^{m} y_i x_i \end{array} \right]. \tag{2.12}$$

Resolvendo este sistema linear, obtemos os valores de $\widehat{a}_0$ e $\widehat{a}_1$ e, logo, a recta dos mínimos quadrados (ou recta de regressão).

No caso da determinação do polinómio dos mínimos quadrados de grau n, o ponto $(\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_n)$ que minimiza a função Φ, é tal que

$$\left\{ \begin{array}{rcl} \dfrac{\partial \Phi}{\partial a_0}(\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_n) & = & 0 \\ & \vdots & \\ \dfrac{\partial \Phi}{\partial a_n}(\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_n) & = & 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} \displaystyle\sum_{i=0}^{m} (y_i - \widehat{a}_0 - \widehat{a}_1 x_i, \cdots - \widehat{a}_n x_i^n) & = & 0 \\ & \vdots & \\ \displaystyle\sum_{i=0}^{m} (y_i - \widehat{a}_0 - \widehat{a}_1 x_i, \cdots - \widehat{a}_n x_i^n) x_i^n & = & 0 \end{array} \right.$$

Temos então um sistema linear para resolver da forma (2.11).

Exercício 2.10 Mostre que o sistema de equações normais (2.12) pode ser escrito na forma $A^T A\widehat{x} = A^T b$, com

$$A = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}, \ \widehat{x} = \begin{bmatrix} \widehat{a}_0 \\ \widehat{a}_1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

Na prática, com o intuito de simplificar a notação, é usual omitir a notação $\widehat{x}$ para representar a solução do sistema $Ax = b$ no sentido dos mínimos quadrados. Assim, dado um sistema de equações lineares $Ax = b$, iremos considerar a solução dos mínimos quadrados (caso exista) como sendo o vector x que é solução do sistema de equações normais $A^T Ax = A^T b$.

Exercício 2.11 Foi efectuado um teste mecânico para estabelecer a relação entre tensões e deformações relativas numa amostra de tecido biológico, tendo-se obtido a seguinte tabela:

tensão σ (N/cm^2)	0,06	0,14	0,25	0,31	0,47	0,60
deformação ϵ (cm)	0,08	0,14	0,20	0,23	0,25	0,28

.

Usando a recta dos mínimos quadrados, obtenha uma estimativa para a deformação correspondente a uma tensão de $\sigma = 0{,}08$ N/cm^2.

Resolução: O sistema de equações normais que perminte obter a recta de regressão $\epsilon = a_0 + a_1\sigma$ pode ser escrito na forma $A^T Ax = A^T b$, com

$$A = \begin{bmatrix} 1 & 0{,}06 \\ 1 & 0{,}14 \\ 1 & 0{,}25 \\ 1 & 0{,}31 \\ 1 & 0{,}47 \\ 1 & 0{,}60 \end{bmatrix}, \ x = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 0{,}08 \\ 0{,}14 \\ 0{,}20 \\ 0{,}23 \\ 0{,}25 \\ 0{,}28 \end{bmatrix},$$

ou seja, é dado por

$$A^T Ax = A^T b \Leftrightarrow \begin{bmatrix} 6 & 1{,}83 \\ 1{,}83 & 0{,}7627 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 1{,}18 \\ 0{,}4312 \end{bmatrix},$$

cuja solução é

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0{,}09035 \\ 0{,}34857 \end{bmatrix}.$$

Concluímos então que a recta de regressão é dada por

$$\epsilon = 0{,}09035 + 0{,}34857\sigma.$$

Uma estimativa para a deformação correspondente a uma tensão de $\sigma = 0{,}08$ N/cm^2 pode ser agora dada por

$$\epsilon = 0{,}09035 + 0{,}34857{,}08 = 0{,}1182356 \text{ cm} \approx 0{,}12 \text{ cm}.$$

Como vimos, a solução dos mínimos quadrados consiste em resolver

$$A^T Ax = A^T b.$$

Este sistema é não singular se A tiver característica máxima n. Se assim for, a matriz $A^T A$ é SPD e a solução dos mínimos quadrados é dada de forma única por

$$x = (A^T A)^{-1} A^T b.$$

A matriz

$$A^+ = (A^T A)^{-1} A^T$$

é chamada a matriz pseudo-inversa da matriz A.

Note-se que, se a matriz A for mal condicionada, a matriz $A^T A$ pode ser muito mal condicionada. Atente-se, por exemplo, ao seguinte exemplo. Seja

$$A = \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix}, \qquad 0 < \epsilon \ll 1.$$

Considerando a norma $\|\cdot\|_2$ temos que $K_2(A) = \rho(A)\rho(A^{-1}) = 1/\epsilon \gg 1$. Assim,

$$K_2(A^T A) = \epsilon^{-2} = \epsilon^{-1} K_2(A) \gg K_2(A).$$

Para contornar este problema, considera-se uma técnica alternativa para resolver o problema dos mínimos quadrados. Essa técnica, mais robusta, consiste em obter a chamada factorização QR da matriz A. Seja $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, com $m \leq n$, uma matriz de característica n. Neste caso, a matriz A pode ser escrita na forma $A = QR$, sendo $Q \in \mathcal{M}_m(\mathbb{R})$ e $R \in \mathcal{M}_{m\times n}(\mathbb{R})$, com Q uma matriz ortogonal, isto é, uma matriz tal que $Q^T Q = QQ^T = I$, e R uma matriz trapezoidal superior cujas linhas são nulas a partir da $(n+1)$-ésima. Assim

$$A^T Ax = A^T b \Leftrightarrow R^T Q^T QRx = R^T Q^T b \Leftrightarrow R^T Rx = R^T Q^T b.$$

A solução dos mínimos quadrados é então obtida na forma

$$x = \hat{R}^{-1} \hat{Q}^T b,$$

onde $\hat{R} \in \mathcal{M}_n(\mathbb{R})$ e $\hat{Q} \in \mathcal{M}_{m\times n}(\mathbb{R})$ são dadas, respectivamente, pelas primeiras n linhas da matriz R e as primeiras n colunas da matriz Q.

2.7 Problemas

2.7.1 Exercícios para resolver nas aulas

Exercício 2.12 Considere o sistema $Ax = b$ onde $A = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}$ e $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

1. Mostre que A admite uma decomposição $A = LU$.
2. Determine a solução dos sistema tendo em atenção a alínea anterior.

Exercício 2.13 Determine a solução do sistema

$$\begin{bmatrix} 1 & 0 & 3 \\ 3 & 6 & 4 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

usando a fatorização LU da matriz do sistema,

1. sem escolha parcial de *pivot*,
2. com escolha parcial de *pivot*.

Exercício 2.14 (Matlab) Determine a factorização LU das matrizes

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{bmatrix} \quad \text{e} \quad B = \begin{bmatrix} 1 & 1 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{bmatrix}.$$

Exercício 2.15 (Matlab) Verifique que a matriz $\mathrm{A} = \begin{bmatrix} 4 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix}$ é simétrica e positiva definida e determine a factorização LL^T de A.

Exercício 2.16 (Matlab) Recorra à factorização LU para resolver o sistema

$$\begin{cases} x_1 + 2x_2 + x_3 + x_4 = 2 \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 6 \\ x_1 + 2x_2 + 2x_3 + x_4 = 2 \\ 3x_1 + 7x_2 - x_3 - x_4 = -8 \end{cases}.$$

Exercício 2.17 (Matlab) Determine a factorização $PA = LU$ das seguintes matrizes, recorrendo à escolha parcial de pivot:

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{bmatrix}; \qquad B = \begin{bmatrix} 0 & 2 & 3 \\ 1 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Exercício 2.18 (Matlab) Determine a solução do sistema

$$\begin{cases} x_1 + x_2 - x_3 = 1 \\ 5x_1 + 2x_2 + 2x_3 = -4 \\ 3x_1 + x_2 + x_3 = 1 \end{cases}:$$

1. usando a factorização LU, sem escolha parcial de pivot;
2. usando a factorização LU, com escolha parcial de pivot.

Exercício 2.19 (Matlab) Um engenheiro electrotécnico supervisiona a produção de três tipos de componentes electrónicas. Três tipos de material - metal, plástico e borracha - são necessários para a produção. As quantidades exigidas para produzir cada componente são indicadas na tabela:

componente	metal(g/componente)	plástico(g/componente)	borracha(g/componente)
1	15	0,30	1,0
2	17	0,40	1,2
3	19	0,55	1,5

.

Se diariamente estiverem disponíveis $3{,}89$, $0{,}095$ e $0,,282$ quilogramas de metal, plástico e borracha, respectivamente, quantas componentes podem ser produzidas por dia?

Exercício 2.20 (Matlab) O sistema $Ax = b$ tem solução única. Use a factorização LU para a determinar, sabendo que

$$A = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 4 & 1 \\ 1 & 2 & 2 & 2 \\ 3 & 7 & -1 & -1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 2 \\ 6 \\ 2 \\ -8 \end{bmatrix}.$$

Exercício 2.21 (Matlab) Considere os dados da tabela

Teste	1	2	3	4	5	6	7	8
Tensão	0,00	0,06	0,14	0,25	0,31	0,47	0,60	0,70
Deformação	0,00	0,08	0,14	0,20	0,23	0,25	0,28	0,29

correspondentes aos valores da deformação para diferentes valores da tensão aplicada numa amostra de tecido biológico (um disco invertebral). Determine a equação da recta de regressão, usando processos diferentes: 1. a instrução `polyfit`; 2. o comando \. Estime o valor da deformação correspondente a uma tensão igual a 0,9.

Exercício 2.22 (Matlab) O seguinte sistema de equações foi obtido aplicando a lei da corrente em rede a um determinado circuito.

$$\left\{\begin{array}{rcrcrcr} 55I_1 & & & - & 25I_4 & = & -200 \\ & - & 37I_3 & - & 4I_4 & = & -250 \\ -25I_1 & - & 4I_3 & + & 29I_4 & = & 100 \end{array}\right. .$$

Resolva o sistema.

Exercício 2.23 Dada a matriz $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & -2 & 0 \\ -2 & -1 & 0 \end{bmatrix}$, calcule $\|A\|_1$, $\|A\|_\infty$ e $\|A\|_2$.

Exercício 2.24 Seja A uma matriz quadrada de ordem n. Mostre que:

1. se A é não singular $K(A) \geq 1$;
2. para a matriz identidade, $K(I) = 1$;
3. para todo o escalar α, $K(\alpha A) = K(A)$;
4. se A é uma matriz diagonal $A = \text{diag}(a_i)$, $K_2(A) = \frac{\max |a_i|}{\min |a_i|}$.

Exercício 2.25 Seja A uma matriz real, não singular e de ordem n. Prove que se λ é um valor próprio de A então

$$\frac{1}{\|A^{-1}\|} \leq |\lambda| \leq \|A\|.$$

Exercício 2.26 Seja A uma matriz simétrica e definida positiva.

1. Mostre que $K_2(A) = \lambda_M/\lambda_m$, onde λ_M e λ_m são, respectivamente, o maior e menor valor próprio de A em valor absoluto.
2. Mostre que $K_2(A^2) = (K_2(A))^2$.

Exercício 2.27 As matrizes dos sistemas

$$\left\{\begin{array}{rcl} x - y & = & 1 \\ x - 1{,}00001y & = & 0 \end{array}\right. \quad \text{e} \quad \left\{\begin{array}{rcl} x - y & = & 1 \\ x - 0{,}99999y & = & 0 \end{array}\right.$$

são aproximadamente iguais. Determine e compare as suas soluções. Explique os resultados obtidos.

Exercício 2.28 Resolva o sistema

$$\left\{\begin{array}{rcl} 2{,}000112x_1 + 1{,}414215x_2 & = & 0{,}521471 \\ 1{,}414215x_1 + 1{,}000105x_2 & = & 0{,}232279 \end{array}\right.$$

pelo método de eliminação de Gauss. Sabendo que $(x_1, x_2) = (607{,}1248,\ -858{,}2826)$ é a sua solução exacta, explique os resultados obtidos.

Exercício 2.29 Seja A a matriz definida por $A = \begin{bmatrix} 1 & a \\ 0 & 2 \end{bmatrix}$, com inversa $A^{-1} = \begin{bmatrix} 1 & -a/2 \\ 0 & 1/2 \end{bmatrix}$.

1. Calcule o número de condição da matriz A associado à norma $\|\cdot\|_1$.
2. Suponha que, ao resolver o sistema $Ax = b$ por eliminação de Gauss, com $a = 10$, encontra uma solução $\widehat{x}$ que satisfaz $\|A\widehat{x} - b\|_1/\|b\|_1 < 10^{-3}$. Determine um majorante para o erro relativo de $\widehat{x}$.

Exercício 2.30 (Matlab) Seja $A_\varepsilon = \begin{bmatrix} \epsilon & 1 \\ 0 & 1 \end{bmatrix}, \qquad 0 < \epsilon \ll 1.$

1. Calcule $K_2(A_\varepsilon)$ e $K_2(A_\varepsilon^T A_\varepsilon)$, quando $\epsilon \to 0$ e represente os resultados graficamente.
2. Suponha que quer resolver o sistema $A_\varepsilon x = b_\varepsilon$, onde b_ε é determinado tal que $x = [1, 1]^T$ e a solução do sistema. Determine como evolui o erro relativo da solução numérica calculada $\bar{x}$ em função de ϵ.

Exercício 2.31 (Matlab) Considere o sistema linear $A_\varepsilon x = b_\varepsilon$, com

$$A_\varepsilon = \begin{bmatrix} 2 & -2 & 0 \\ \varepsilon - 2 & 2 & 0 \\ 0 & -1 & 3 \end{bmatrix}$$

e b_ε é escolhido por forma a que a solução do sistema seja $x = [1, 1, 1]^T$, sendo ε um número real positivo.

1. Seja $\varepsilon = 0{,}1$. Calcule a factorização LU com e sem escolha parcial de pivot. Determine a solução do sistema linear usando ambas as factorizações. Calcule $\|A_\varepsilon \widehat{x} - b_\varepsilon\|_2$, sendo $\widehat{x}$ cada uma das soluções obtidas.
2. Mostre experimentalmente que se $\varepsilon \to 0$ então $|l_{32}| \to \infty$ (sem efectuar escolha parcial de pivot). O que acontece ao número de condição da matriz A_ε e dos factores L e U (relativamente à norma $\|\cdot\|_2$) à medida que $\varepsilon \to 0$?.
3. Considerando novamente $\varepsilon \to 0$, qual o comportamento do erro relativo do sistema em função de ε? Comente os resultados.

Exercício 2.32 (Matlab) Considere a matriz $H_n = (h_{ij})_{i,j=1}^n$ tal que

$$h_{ij} = \frac{1}{i + j - 1}, \qquad i, j = 1, ..., n,$$

chamada matriz de Hilbert de ordem n, em homenagem ao famoso matemático alemão David Hilbert (1862-1943).

1. Fazendo n variar entre 2 e 10, determine o número de condição (relativamente à norma $\|\cdot\|_2$) da matriz H_n e represente os resultados graficamente.
2. Supondo que se pretende resolver o sistema $H_n x = b$, onde b é determinado tal que $x = [1, 1, ..., 1]^T$ é a solução do sistema, determine como evolui o erro relativo da solução numérica calculada $\bar{x}$ em função de n (faça variar n entre 2 e 10).

Exercício 2.33 Considere $A = \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ -a & 0 & 1 \end{bmatrix}$, com $a \in \mathbb{R}$, e verifique que

$$A^{-1} = \begin{bmatrix} 1/(1+a^2) & 0 & -a/(1+a^2) \\ 0 & 1 & 0 \\ a/(1+a^2) & 0 & 1/(1+a^2) \end{bmatrix}.$$

1. Calcule as normas $\|\cdot\|_\infty$ e $\|\cdot\|_1$ da matriz A.
2. Calcule $K_\infty(A)$ e $K_1(A)$. Para que valores de a há mau condicionamento da matriz?

Exercício 2.34 Aplicando o método de Jacobi, determine uma aproximação da solução do seguinte sistema

$$\begin{bmatrix} 3 & -1 & 1 \\ 3 & 6 & 2 \\ 3 & 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix},$$

começando com uma aproximação inicial $x^{(0)} = [0, 0, 0]^T$.

Exercício 2.35 Obtenha duas aproximações para a solução do seguinte sistema linear

$$\begin{cases} -8x + y + z = 1 \\ x - 5y + z = 16 \\ x + y - 4z = 7 \end{cases},$$

partindo do vector inicial $x^{(0)} = [0, 0, 0]^T$, usando o método de Jacobi e o método de Gauss-Seidel. Qual o erro absoluto e relativo das aproximações obtidas?

Exercício 2.36 Considere o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 4/3 & 2 & 1/2 \\ 0 & 1 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 4 \\ 6 \\ 4 \end{bmatrix}.$$

1. Deduza a expressão do método de Jacobi aplicado à resolução do sistema linear $Ax = b$.
2. Mostre que o método de Jacobi converge para a solução do sistema, qualquer que seja a aproximação inicial escolhida.
3. Efetue duas iterações do método de Jacobi considerando a aproximação inicial $x^{(0)} = [0; 0; 0]^T$.
4. Determine quantas iterações do método de Jacobi devem ser efetuadas para garantir que o erro inicial, medido na norma $\|\cdot\|_\infty$, seja reduzido de um factor de 10^{-7}.

Exercício 2.37 Considere o sistema linear

$$\begin{cases} x - 2y = -2 \\ 2x + y = 2 \end{cases}.$$

1. Verifique que o método de Gauss-Seidel aplicado ao sistema diverge.
2. Reordene as equações de modo a obter um sistema equivalente que lhe permita garantir que este método converge.

Exercício 2.38 Para aproximar a solução (x_1, x_2, x_3) de um sistema linear $Ax = b$, recorreu-se ao seguinte método iterativo

$$\begin{cases} x_1^{(k+1)} &= -0{,}6x_2^{(k)} - 0{,}6x_3^{(k)} + 1 \\ x_2^{(k+1)} &= -0{,}6x_1^{(k)} - 0{,}6x_3^{(k)} + 1 \\ x_3^{(k+1)} &= -0{,}6x_1^{(k)} - 0{,}6x_2^{(k)} + 1 \end{cases}, \qquad k = 0, 1, \ldots.$$

1. Escreva a respectiva matriz de iteração. O método será convergente para todo o ponto inicial?
2. O método apresentado pode ser identificado com o método de Jacobi ou com o método de Gauss-Seidel? Justifique a sua resposta.
3. Sabendo que $b = [1, 1, 1]^T$, obtenha a matriz A.

Exercício 2.39 Considere o sistema linear $\begin{cases} 4x - y - z = 2 \\ x + ky + 3z = 4 \\ x + 2y + 0{,}5z = 4 \end{cases}$.

1. Determine os valores do parâmetro k para os quais o sistema tem uma só solução.
2. Para $k = 0$ poderá aplicar o método de Gauss-Seidel sem alterar o sistema? Justifique.
3. Determine valores de k para os quais seja garantida a convergência do método de Jacobi.
4. Faça $k = 0$ e calcule duas aproximações para a solução do sistema, utilizando o método de Jacobi.

Exercício 2.40 Considere o sistema linear $\begin{cases} x - y - z = -1 \\ 2y + az = 0 \\ -x + 2z = 3 \end{cases}$, com $a \in \mathbb{R}^-$.

1. Determine todos os valores do parâmetro a que garantem a convergência do método de Gauss-Seidel quando aplicado a este sistema.
2. Para $a = -1$ efectue duas iterações do referido método, indicando uma estimativa para o erro cometido.

Exercício 2.41 Considere o sistema

$$\begin{bmatrix} 5 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 5 \end{bmatrix} x = \begin{bmatrix} 6 \\ 7 \\ 0 \end{bmatrix}.$$

1. Prove que o polinómio característico associado à matriz de iteração do método de Gauss-Seidel, quando aplicado ao sistema anterior, é $P(\lambda) = -\lambda^3 + \frac{46}{75}\lambda^2 - \frac{2}{25}\lambda$.
2. Localize e separe as raízes de $P(\lambda) = 0$.
3. O método de Gauss-Seidel, aplicado ao sistema anterior, é convergente? Justifique.
4. Determine a segunda aproximação gerada pelo método de Gauss-Seidel, quando aplicado ao sistema anterior.

Exercício 2.42 Considere a matriz $A = \begin{bmatrix} 0 & 2 & 1 \\ -1 & 1 & 2 \\ 0 & -3 & -1 \end{bmatrix}$.

1. Mostre que o polinómio característico associado a A é $P(\lambda) = -\lambda^3 - 7\lambda + 1$.
2. Localize e separe todos os valores próprios de A.
3. Seja A a matriz de iteração de um método iterativo que aproxima a solução de um sistema de equações lineares $Cx = d$. Será que, recorrendo ao resultado da alínea anterior, pode tirar alguma conclusão acerca da convergência desse método iterativo? Justifique.

Exercício 2.43 (Matlab) O sistema $\begin{cases} 5x - y = 3 \\ -x + 10y = 19 \end{cases}$ tem a solução $[1, 2]^T$. Aproxime-a usando os métodos iterativos de Jacobi e Gauss-Seidel com $x^{(0)} = [0, 0]^T$, e compare os resultados.

Exercício 2.44 (Matlab) Aplique os métodos de Jacobi e Gauss-Seidel para aproximar a solução do sistema

$$\begin{bmatrix} \alpha & 0 & 1 \\ 0 & \alpha & 0 \\ 1 & 0 & \alpha \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

para $\alpha = 2$ e $\alpha = -2$. Comente os resultados obtidos.

Exercício 2.45 (Matlab) Verifique se os métodos de Jacobi e Gauss-Seidel convergem quando aplicados aos seguintes sistemas:

1. $\begin{cases} 9x + 3y + z = 13 \\ -6x + 8z = 2 \\ 2x + 5y - z = 6 \end{cases}$; 2. $\begin{cases} x + y + 6z = 8 \\ x + 5y - z = 5 \\ 4x + 2y - 2z = 4 \end{cases}$;

3. $\begin{cases} -3x + 4y + 5z = 6 \\ -2x + 2y - 3z = -3 \\ 2y - z = 1 \end{cases}$.

Exercício 2.46 (Matlab) Uma fábrica de equipamento electrónico produz transistores, resistências e chips de computadores. Para a respectiva construção, os materiais exigidos são cobre, zinco e vidro. O número de unidades necessárias para cada componente são indicadas na tabela:

componente	cobre	zinco	vidro
transistores	4	1	2
resistências	3	3	1
chips de computadores	2	1	3

.

Numa determinada semana as quantidades de materiais disponíveis são 960 unidades de cobre, 510 de zinco e 610 de vidro.

1. Obtenha o sistema que permite determinar o número de componentes de cada tipo que podem ser produzidas naquela semana.
2. Aproxime a solução do sistema recorrendo aos métodos de Jacobi e de Gauss-Seidel.
3. Compare os resultados obtidos em na alínea anterior com a solução exacta.

Exercício 2.47 Considere o sistema $A_\alpha x = b$ onde

$$A_\alpha = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 2 & \alpha \\ -1 & 0 & 2 \end{bmatrix} \text{ e } b = \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix}$$

1. Determine todos os valores de α para os quais o método de Gauss-Seidel é convergente.
2. Para $\alpha = -1$, efetue duas iterações do referido método.

Exercício 2.48 Estabelaça uma condição suficiente para β de tal modo que os métodos de Jacobi e Gauss-Seidel sejam convergentes quando aplicados a aproximar a solução de um sistema com matriz

$$\begin{bmatrix} -10 & 2 \\ \beta & 5 \end{bmatrix}.$$

Exercício 2.49 Considere o sistema linear

$$\begin{cases} 10x - 9y = -1 \\ -x + 31y = 2 \end{cases}.$$

1. Escreva o método de Gauss-Seidel na forma matricial para o sistema dado.
2. Pode garantir a convergência do método? Justifique.
3. Aplicando (duas vezes) o método de Gauss-Seidel, determine uma aproximação para a solução do sistema partindo da aproximação inicial $x^{(0)} = [0; 0]^T$.

Exercício 2.50 Considere o sistema linear $Ax = b$, onde

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

1. Escreva o método de Jabobi na forma matricial para o sistema dado.
2. Pode garantir a convergência do método? Justifique.
3. Aplicando (duas vezes) o método de Jacobi, determine uma aproximação para a solução do sistema partindo da aproximação inicial $x^{(0)} = [0\ 0\ 0]^T$. Calcule o fator de redução do erro absoluto associado à segunda iteração do método de Jacobi, face ao erro inicial.

Exercício 2.51 Determine a linha recta que melhor se ajusta, no sentido dos mínimos quadrados, aos seguintes pontos (e represente graficamente):

1. $(-1, 2),\ (1, -3),\ (2, -5),\ (0, 0)$; 2. $(1, 1),\ (2, 5),\ (3, 7),\ (4, 9),\ (5, 12)$.

Exercício 2.52 Determine a solução no sentido dos mínimos quadrados do sistema (com m equações e uma incógnita) $x = \beta_1,\ x = \beta_2, \ldots,\ x = \beta_m$.

Exercício 2.53 O sistema

$$\begin{cases} x_1 + 2x_2 = 1 \\ 2x_1 + 5x_2 = 0 \\ 3x_1 + 7x_2 = 2 \end{cases}$$

é impossível. Verifique que a solução no sentido dos mínimos quadrados é única.

Exercício 2.54 O proprietário de uma empresa em rápido crescimento económico verificou que, nos primeiros seis anos, o lucro, L, da sua empresa em função do número de anos decorridos, N, poderia ser aproximado por uma transformação linear $L = a + bN$. Atendendo a que os resultados do seu negócio foram

N (número de anos)	0	1	3	6
L (lucro, em milhares de euros)	0	1	3	4

,

determine:

1. a recta dos mínimos quadrados para o problema descrito;
2. um valor para o lucro previsível no final do sétimo ano.

Exercício 2.55 Verificar que a recta de regressão passa pelo ponto cuja abcissa é a média dos x_i, $i = 1, \ldots, n$, e cuja ordenada é a média dos $f(x_i)$, $i = 1, \ldots, n$.

Exercício 2.56 (Matlab) Calcule a parábola dos mínimos quadrados para a função f dada pela seguinte tabela

x_i	0	0,1	0,2	0,3	0,4	0,5	0,6
$f(x_i)$	2,9	2,8	2,7	2,3	2,1	2,1	1,7

.

Exercício 2.57 Determine a função da forma $f(x) = ax^2 + b\sqrt{x}$ que melhor se ajusta, no sentido dos mínimos quadrados, aos dados da tabela seguinte

x	0	1	3	4	5
y	10	50	71	95	122

.

Exercício 2.58 Considere os pontos $(0, 1)$, $(1, 2)$, $(2, 5)$, $(4, 8)$. Determine as constantes a e b por forma a que a função

$$f(x) = a(x + 1) + bx^2$$

se ajuste aos dados no sentido dos mínimos quadrados.

Exercício 2.59 Determine as constantes a e b por forma a que $y = be^{ax}$ se ajuste aos dados da tabela

x_i	1	1,25	1,5	1,75	2,0
y_i	5,1	5,8	6,5	7,5	8,4

no sentido dos mínimos quadrados.

Exercício 2.60 A pressão sistólica p (em milímetros de mercúrio) de uma criança saudável com peso w (em quilogramas) é dada, de forma aproximada, pela equação $p = a + b \ln w$. Use os seguintes dados experimentais

w	20	28	37	51	59
p	91	99	104	108	111

para estimar a pressão sistólica de uma criança de 45 quilogramas.

Exercício 2.61 Pretende-se ajustar a função

$$g(x) = a + b \sin x, \qquad a, b \in \mathbb{R},$$

no sentido dos mínimos quadrados, aos pontos da seguinte tabela:

x_i	0,0	1,0	2,0	3,0	4,0
f_i	1,0	1,8	1,9	1,1	0,2

1. Mostre que a e b verificam $Mx = d$, com M uma matriz a determinar, $x = [a, b]^T$ e

$$d = \begin{bmatrix} 6{,}0 \\ 3{,}2462 \end{bmatrix}.$$

2. Suponhamos que se pretende resolver o sistema linear anterior com o segundo membro dado por

$$\bar{d} = \begin{bmatrix} 6{,}00 \\ 3{,}2 \end{bmatrix}.$$

Mostre que a solução $\bar{x}$ de $M\bar{x} = \bar{d}$ verifica

$$\frac{\|x - \bar{x}\|}{\|x\|} \leq K(M) \frac{\|d - \bar{d}\|}{\|d\|}$$

e deduza qual o erro relativo que se comete quando se considera a norma $\|\cdot\|_\infty$ e a aproximação dada para o segundo membro.

Exercício 2.62 Pretende-se ajustar a função

$$g(x) = a + bx^2, \qquad a, b \in \mathbb{R},$$

no sentido dos mínimos quadrados, aos pontos da seguinte tabela:

x_i	0,0	1,5	3,0	4,5	6,0
f_i	1,00	1,57	2,00	4,30	7,00

1. Mostre que a e b verificam

$$Mx = d \Leftrightarrow \begin{bmatrix} 5 & 67{,}5 \\ 67{,}5 & 1792{,}125 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 15{,}87 \\ 360{,}6075 \end{bmatrix}$$

2. A resolução do sistema anterior pode ser feita recorrendo ao método de Gauss-Seidel. Mostre que este método é convergente quando aplicado ao problema anterior.

3. Mostre que, para um método iterativo $x^{(k+1)} = Bx^{(k)} + c$, $k = 0, 1, ...$, usado na aproximação de x^* se tem

$$\|x^* - x^{(k)}\| \leq \|B\|^k \|x^* - x^{(0)}\|$$

e determine quantas iterações são necessárias do método de Gauss-Seidel por forma a reduzir o erro inicial da aproximação por um fator de 10^{-3}.

Exercício 2.63 (Matlab) Seja

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}.$$

Calcule a aproximação dos mínimos quadrados para o problema $Ax = b$ usando processos diferentes: 1. a instrução `polyfit`; 2. o comando \; 3. a instrução `pinv`; 4. a decomposição `qr`.

Exercício 2.64 A lei de Hooke, devida a Robert Hooke (1635-1703), estabelece que a força F aplicada a uma mola é directamente proporcional ao deslocamento provocado de acordo com a seguinte relação

$$F = k(e - e_0),$$

onde k é a constante da mola, e o comprimento da mola quando sujeita à força F e e_0 o comprimento inicial da mola.

No sentido de determinar a constante da mola usaram-se diferentes forças (conhecidas) tendo sido observados os comprimentos resultantes, dados na seguinte tabela

força F (em gramas)	3	5	8	10
comprimento e (em milímetros)	13,3	16,3	19,4	20,9

.

Sabendo que o comprimento inicial da mola é $e_0 = 10$ mm, determine a melhor estimativa para a constante da mola no sentido dos mínimos quadrados.

2.7.2 Exercícios de aplicação à engenharia

Exercício 2.65 Considere um barra horizontal fixa numa extremidade e livre no restante do seu comprimento. Um modelo discreto de forças na barra conduz ao sistema de equações lineares $Ax = b$, onde A é a matriz quadrada de ordem n, com estrutura de banda, dada por

$$\begin{bmatrix} 6 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & 1 \\ & & & & 1 & -4 & 6 \end{bmatrix}.$$

O vector b é dado pela carga que é imposta à barra (incluindo o seu próprio peso), e o vector x representa a deformação da barra que queremos determinar. Considermos a barra sujeita a uma carga uniforme dada por $b_i = 1$, para $i = 1, \ldots, n$. Considerando $n = 100$, resolva o sistema usando métodos directos e métodos iterativos, comparando a sua eficácia.

Capítulo 3

Valores próprios e valores singulares

Dada uma matriz $A \in \mathcal{M}_n(\mathbb{C})$, os valores próprios de A são os escalares λ para os quais

$$Ax = \lambda x, \qquad x \in \mathbb{C}^n, \qquad x \neq 0.$$

A x chama-se vector próprio associado ao valor próprio λ.

Note-se que o vector próprio associado a um determinado valor póprio não é dado de forma única. Se x é vector próprio de A, então αx também é, com α um escalar não nulo.

Conhecendo um vector próprio x de uma matriz A, o valor próprio associado é dado pelo quociente de Rayleigh,

$$\lambda = \frac{\overline{x}^T A x}{\|x\|^2},$$

com $\overline{x}^T$, isto é o vector transposto do seu conjugado. Note-se que, se $x \in \mathbb{R}^n$, $\overline{x}^T = x^T$. Este quociente é assim chamado como homenagem Lord Rayleigh (1842-1919).

A forma usada nas disciplinas de Álgebra Linear para determinar os valores próprios de uma matriz, é a que recorre ao uso do polinómio característico, obtido sucessivamente por

$$Ax = \lambda x, \Leftrightarrow (A - \lambda I)x = 0 \Rightarrow \det(A - \lambda I) = 0,$$

uma vez que $x \neq 0$. Os valores próprios de A são as n raízes, reais com complexas, iguais ou distintas, do polinómio característico $\det(A - \lambda I) = 0$, em λ, de grau n.

Por outro lado, dado um polinómio mónico arbitrário

$$p(\lambda) = \lambda^n + c_{n-1}\lambda_{n-1} + \cdots + c_1\lambda + c_0,$$

ele é o polinómio característico da sua matriz companheira

$$C_n = \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}.$$

Como as raízes de um poliómio de grau superior a 4 não podem, em geral, ser calculadas num número finito de passos temos que o mesmo se passa para o cálculo de valores próprios de matrizes com ordem superior a 4. Temos, pois, que pensar em métodos iterativos para o cálculo dos valores próprios de uma matriz.

3.1 Método da potência

Considere-se $A \in \mathcal{M}_n(\mathbb{C})$ uma matriz com valores próprios λ_i, $i = 1, ..., n$, tais que

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Seja x_i o vector próprio de módulo 1 associado a λ_i, isto é, tal que $\|x_i\| = 1$. Se os vectores próprios de A forem linearmente independentes, o método da potência, que calcula a sucessão de vectores $x^{(k+1)} = Ax^{(k)}$, $k = 0, 1, 2, ...$, partindo de um vector não nulo $x^{(0)}$, permite obter o vector próprio x_1 associado ao maior valor próprio (em módulo) λ_1 da matriz A.

De facto, considere-se $x^{(0)}$ escrito como combinação linear dos vectores própios de A, isto é,

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i x_i.$$

Tal combinação é sempre possível pois assumimos que os vectores próprios de A são linearmente independentes. Então

$$x^{(k)} = Ax^{(k-1)} = \cdots = A^k x^{(0)} = \sum_{i=1}^{n} \lambda_i^k \alpha_i x_i.$$

Colocando λ_1^k em evidência ($\lambda_1 \neq 0$), temos que

$$x^{(k)} = \lambda_1^k \left(\alpha_1 x_1 + \sum_{i=2}^{n} \left(\frac{\lambda_i}{\lambda_1} \right)^k \alpha_i x_i \right).$$

Como $\frac{\lambda_i}{\lambda_1} < 1$, para todo o $i > 1$, temos que a sucessão converge para um múltiplo de x_1.

Uma vez calculado o vector próprio x_1, por normalização, o valor próprio λ_1 é calculado usando o quociente de Rayleigh.

Exemplo 3.1 Considere-se a matriz A e o vector $x^{(0)}$ dados por

$$A = \begin{bmatrix} 1{,}5 & 0{,}5 \\ 0{,}5 & 1{,}5 \end{bmatrix}, \qquad x^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Pelo método da potência obtemos

k	$\left(x^{(k)}\right)^T$		$\lambda_1^{(k)}$
0	0,0	1,0	
1	0,5	1,5	1,8000
2	1,5	2,5	1,9412
3	3,5	4,5	1,9846
4	7,5	8,5	1,9961
5	15,5	16,5	1,9990
6	31,5	32,5	1,9998
7	63,5	64,5	1,9999
8	127,5	128,5	2,0000

que, permite antever a convergência do método para o valor próprio de maior módulo, que é igual a 2.

O crescimento geométrico das componentes do vector em cada iteração põe em risco o eventual overflow (ou underflow se $\lambda_1 < 1$). Para contornar esse problema, as sucessivas aproximações ao vector próprio devem ser normalizadas, dando lugar ao esquema iterativo

$$x^{(k)} = Ay^{(k-1)}, \qquad y^{(k)} = x^{(k)}/\|x^{(k)}\|,$$

a partir de $y^{(0)} = x^{(0)}/\|x^{(0)}\|$. Com esta normalização, obtemos o seguinte algoritmo.

Algoritmo 3.1 Método da potência

Dados: A, $x^{(0)}$, ε
$y := x^{(0)}/\|x^{(0)}\|$
$\lambda := \overline{y}^T Ay$
$k := 0$
$erro := \varepsilon|\lambda|+1$
Enquanto $erro > \varepsilon|\lambda|$ fazer
    $k := k + 1$
    $x := Ay$
    $y := x/\|x\|$
    $\mu := \overline{y}^T Ay$
    $erro := |\mu - \lambda|$
    $\lambda := \mu$
Resultado: $\lambda_1 \approx \lambda$

A velocidade de convergência do método da potência depende da razão $|\lambda_2/\lambda_1|$. Escolhendo um escalar σ tal que

$$\frac{|\lambda_2 - \sigma|}{|\lambda_1 - \sigma|} < \frac{|\lambda_2|}{|\lambda_1|} \tag{3.1}$$

a convergência do método da potência pode ser acelerada.

Exercício 3.1 Mostre que, se λ é um valor próprio de A,

1. $\lambda - \sigma$ é um valor próprio de $A - \sigma I$, com I a matriz identidade da mesma ordem da matriz A;
2. com A uma matriz invertível, então λ^{-1} é um valor próprio de A^{-1}.

Atendendo ao exercício anterior, temos que a convergência do método da potência pode ser acelerada se considerarmos, na aproximação do valor próprio de maior módulo da matriz $A - \sigma I$, um σ que verifique (3.1). Para calcular o valor próprio de maior módulo de A basta adicionar σ ao resultado.

Atendendo ainda ao exercício anterior podemos concluir que o método da potência permite calcular o menor valor próprio (em módulo) de uma matriz A, calculando o maior valor próprio da sua inversa. Neste caso, temos o esquema iterativo

$$Ax^{(k)} = y^{(k-1)}, \qquad y^{(k)} = x^{(k)}/\|x^{(k)}\|,$$

que evita o cálculo explícito da matriz inversa. O método inverso da potência pode ser usado de forma muito eficaz se uma boa aproximação para o valor próprio de menor módulo da matriz A for conhecida. Essa aproximação pode ser obtida pelo quociente de Rayleigh, de acordo com o esquema iterativo

$$\sigma_k = \left(\overline{y^{(k-1)}}\right)^T A y^{(k-1)}, \qquad (A - \sigma_k I)x^{(k)} = y^{(k-1)}, \qquad y^{(k)} = x^{(k)}/\|x^{(k)}\|.$$

Suponhamos agora que pretendemos calcular o valor próprio λ_σ de A que está mais próximo de σ, com σ um número real ou complexo. Neste caso, o menor valor próprio de $A - \sigma I$ é $\lambda_{min} = \lambda_\sigma - \sigma \approx 0$. Para calcular λ_σ procede-se da seguinte forma, cuja justificação se encontra expressa no exercício anterior: calcula-se λ_{min}, o menor valor próprio (em módulo) de $A - \sigma I$, pelo método inverso da potência; faz-se $\lambda_\sigma = \lambda_{min} + \sigma$.

Para ter uma ideia onde se situam os valores próprios de uma matriz $A = [a_{i,j}]_{i,j=1}^n$, podem usar-se os círculos de Gershgorin

$$C_i^{(l)} = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}, \qquad , i = 1, ..., n,$$

ou

$$C_j^{(c)} = \left\{ z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{i=1, i\neq j}^{n} |a_{ij}| \right\}, \qquad j = 1, ..., n,$$

definidos por Semyon Aranovich Gershgorin (1901-1933). A $C_i^{(l)}$ e $C_j^{(c)}$ chamam-se, respectivamente, o i-ésimo círculo por linhas e j-ésimo círculo por colunas.

Demonstra-se que todos os valores próprios de uma matriz $A \in \mathcal{M}_n(\mathbb{C})$ pertencem à região do plano complexo definido pela intersecção das duas regiões constituídas, respectivamente, pela união dos círculos de Gershgorin por linhas e pela união dos círculos por colunas. Além disso, se os m círculos por linhas (ou por colunas), com $1 \leq m \leq n$, forem disjuntos da união dos restantes $m - n$ círculos, então a sua união contém exactamente m valores próprios.

3.2 Cálculo de todos os valores próprios

Definição 3.1 (Matrizes semelhantes) *Duas* matrizes A *e* B *são* semelhantes *se existir uma matriz invertível* P *tal que* $P^{-1}AP = B$.

Pode provar-se facilmente que duas matrizes semelhantes têm os mesmos valores próprios. De facto, se λ é valor próprio de A então $Ax = \lambda x$, com $x \neq 0$. Multiplicando por P^{-1} à esquerda ambos os membros desta iqualdade obtemos

$$BP^{-1}x = \lambda P^{-1}x \Leftrightarrow By = \lambda y,$$

com $y = P^{-1}x$, ou seja, λ é valor próprio de B.

O método que iremos apresentar baseia-se na decomposição QR. Seja $A \in \mathcal{M}_n(\mathbb{R})$. A ideia consiste em constuir uma sucessão de matrizes $A^{(k)}$, todas semelhantes a A, que tende para uma matriz cujos valores próprios são simples de calcular.

Considerem-se $Q^{(0)}$ e $R^{(0)}$ as matrizes obtidas pela decomposição QR da matriz A. Defina-se $A^{(1)} = R^{(0)}Q^{(0)}$ e considerem-se as matrizes $Q^{(1)}$ e $R^{(1)}$ obtidas pela decomposição QR da matriz $A^{(1)}$. Podemos obter assim a matriz $A^{(2)} = R^{(1)}Q^{(1)}$. É possível demonstrar que, se os valores próprios de A forem tais que

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|,$$

$$\lim_{k\to+\infty} A^{(k)} = T = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ & \lambda_2 & \cdots & t_{2n} \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

com

$$A^{(k)} = R^{(k-1)}Q^{(k-1)},$$

sendo $Q^{(k-1)}$ e $R^{(k-1)}$ as matrizes obtidas pela decomposição QR da matriz $A^{(k-1)}$.

3.3 Decomposição em valores singulares

Uma matriz diz-se diagonalizável se for semelhante a uma matriz diagonal. Prova-se facilmente que se $A \in \mathcal{M}_n(\mathbb{C})$ tiver n vectores próprios linearmente independentes, então $U^{-1}AU = \Lambda$, com U a matriz cujas colunas são os vectores próprios de A e Λ a matriz diagonal formada pelos correspondentes valores próprios.

Se A for uma matriz real simétrica a decomposição em valores próprios $A = U\Lambda U^{-1}$ é sempre possível e as colunas de U são os vectores próprios de A que podem ser escolhidos por forma a constituir uma base ortonormada de $\mathbb{R}^n$. Nesse caso, a matriz U é uma matriz ortogonal (unitária no caso complexo) uma vez que $U^TU = UU^T = I$ e a decomposição em valores próprios pode ser escrita na forma $A = U\Lambda U^T$

Consideremos agora uma matriz arbitrária $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ (a generalização para o caso complexo pode ser feita facilmente). Neste caso prova-se que a decomposição em valores singulares

$$A = U\Sigma V^T \tag{3.2}$$

existe sempre, com $U \in \mathcal{M}_m(\mathbb{R})$ e $V \in \mathcal{M}_n(\mathbb{R})$ duas matrizes ortogonais e $\Sigma \in \mathcal{M}_{m\times n}(\mathbb{R})$ uma matriz diagonal de elemento genérico

$$\sigma_{ij} = \begin{cases} 0, & i \neq j, \\ \sigma_i \geq 0, & i = j. \end{cases}$$

Os elementos σ_i da diagonal são chamados valores singulares da matriz A e podem ser ordenados na forma

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p,$$

com $p = min\{m, n\}$. As colunas u_i de U e v_i de V são chamados os vectores singlares esquerdos e direitos, respectivamente.

Exemplo 3.2 Seja

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}.$$

A decomposição em valores singulares de A é dada por

$$U\Sigma V^T = \begin{bmatrix} 0{,}141 & 0{,}825 & -0{,}420 & -0{,}351 \\ 0{,}344 & 0{,}426 & 0{,}298 & 0{,}782 \\ 0{,}547 & 0{,}0278 & 0{,}664 & -0{,}509 \\ 0{,}750 & -0{,}371 & -0{,}542 & 0{,}0790 \end{bmatrix} \begin{bmatrix} 25{,}5 & 0 & 0 \\ 0 & 1{,}29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0{,}505 & 0{,}574 & 0{,}644 \\ -0{,}761 & -0{,}057 & 0{,}646 \\ 0{,}408 & -0{,}816 & 0{,}408 \end{bmatrix}.$$

O seguinte resultado mostra que o número de valores próprios não nulos de uma matriz coincide com a característica dessa matriz.

Teorema 3.1 *Se $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ tiver característica r, então A tem r valores singulares positivos.*

Demonstração: Uma vez que U é quadrada e com característica m e que V^T é quadrada e com característica n, sabemos, pelas propriedades da característica de uma matriz, que

$$\text{car}(A) = \text{car}(U\Sigma V^T) = \text{car}(\Sigma).$$

Mas Σ é uma matriz em escada com r pivots (os valores singulares σ's positivos), o que mostra o pretendido. □

Os vectores singulares à esquerda (as colunas da matriz U) contêm uma base para o espaço das colunas de A. Os vectores singulares à direita (as colunas da matriz V) contêm uma base para o espaço nulo de A. Estes factos são enunciados e provados de seguida.

Teorema 3.2 *Seja $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ uma matriz com característica r. Seja $A = U\Sigma V^T$ a sua decomposição em valores singulares. Então:*

1. *Uma base para $C(A)$ é dada por $u_1, \ldots, u_r$.*
2. *Uma base para $N(A)$ é dada por $v_{r+1}, \ldots v_n$.*

Demonstração: Como $\dim(C(A)) = r$ e $\dim(N(A)) = n - r$ e quer as colunas de U quer as de V são linearmente independentes, basta provar que $u_i \in C(A)$, $i = 1, \ldots, r$, e que $v_i \in N(A)$, $i = r+1, \ldots, n$.

É fácil verificar que $u_i \in C(A)$, para um dado $i \in \{1, \ldots, r\}$, multiplicando A por v_i:

$$Av_i = U\Sigma V^T v_i = Ue_i = \sigma_i U e_i = \sigma_i u_i,$$

em que e_i designa a i-ésima coluna da matriz identidade de ordem n. Como $Av_i \in C(A)$, tem-se que $u_i = (1/\sigma_i)Av_i \in C(A)$.

Para mostrar que $v_i \in N(A)$, com $i \in \{r+1, \ldots, n\}$, calcula-se, novamente, o mesmo produto Av_i, mas desta vez com índices i correspondentes às colunas nulas de Σ, obtendo-se

$$Av_i = U\Sigma V^T v_i = U\Sigma e_i = 0. \quad \square$$

Notemos a relação que existe entre a decomposição em valores singulares e a decomposição em valores próprios. Consideremos a matriz simétrica A^TA que tem todos os seus valores próprios são reais. Temos que

$$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T(U^TU)\Sigma V^T.$$

Como U é ortogonal, $U^TU = I$ e então

$$A^TA = V\Sigma^2 V^T.$$

A matriz $\Sigma^2 = \Sigma^T\Sigma \in \mathcal{M}_n(\mathbb{R})$. Observamos, assim, que A^TA e Σ^2 são semelhantes (pois $V^T = V^{-1}$) e, como tal, têm os mesmos valores próprios. Logo, se A tiver característica r, os valores próprios não nulos de A^TA são $\sigma_1, \ldots, \sigma_r$. Provámos, assim, o seguinte teorema.

Teorema 3.3 *Seja $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ uma matriz com característica r. Então, os r valores singulares não nulos de A são as raízes quadradas dos r valores próprios de A^TA.*

Exemplo 3.3 Pretende-se calcular a decomposição em valores singulares de

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

Para isso, calculam-se os valores próprios de

$$A^TA = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix},$$

que facilmente se prova serem iguais a 8 e 2, e os respectivos vectores próprios (ortonormados)

$$v_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{e} \quad v_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

Temos então

$$\Sigma = \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} = \begin{bmatrix} 2\sqrt{2} & \\ & \sqrt{2} \end{bmatrix} \quad \text{e} \quad V = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

As colunas da matriz U são obtidas fazendo

$$\sigma_1 u_1 = Av_1 = \begin{bmatrix} 2 & 2 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix} \Rightarrow u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

e

$$\sigma_2 u_2 = Av_2 = \begin{bmatrix} 2 & 2 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix} \Rightarrow u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Assim, a decomposição em valores singulares de $A = U\Sigma V^T$ é

$$\begin{bmatrix} 2 & 2 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2\sqrt{2} & \\ & \sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Exemplo 3.4 Calculemos, agora, a decomposição em valores singlares de

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}.$$

Como

$$AA^T = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix},$$

tem-se que

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Para construir V basta encontrar um conjunto ortonormado de vectores, próprios da matriz AA^T, associados a cada um dos valores singulares não nulos. Assim

$$V = [v_1 \quad v_2] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Para determinar a matriz ortogonal U, temos que as duas primeiras colunas são obtidas fazendo

$$u_i = \frac{1}{\sigma_i} A v_i,$$

ou seja

$$u_1 = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad u_2 = \frac{1}{1}\begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

O vector u_3 será determinado por forma a ser ortogonal a $\{u_1, u_2\}$ usando o processo de ortogonalização de Gram-Schmidt. Obtemos assim

$$u_3 = \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

e então

$$A = U\Sigma V^T = \begin{bmatrix} 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Existem muitas aplicações práticas onde o conhecimento da decomposição em valores singulares de uma matriz pode ser importante. Uma delas prende-se com o cálculo da norma de uma matriz A. Atendendo ao teorema anterior, tem-se que $\|A\|_2 = \sigma_1$. Como consequência, temos que

$$K_2(A) = \frac{\sigma_1}{\sigma_r}.$$

Também se pode provar que se A tiver característica r, $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$.

No capítulo anterior definimos a matriz pseudo-inversa de uma matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ com $m \geq n$, de característica máxima n, como sendo

$$A^+ = (A^T A)^{-1} A^T.$$

A noção de matriz pseudo-inversa pode ser generalizada para qualquer matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ (mesmo não sendo $A^T A$ invertível), a partir da decomposição em valores singulares. Se $A = U\Sigma V^T$, então

$$A^+ = V\Sigma^+ U^T,$$

onde Σ^+ é obtida transpondo Σ e invertendo as suas entradas não nulas.

Outro exemplo prende-se com a solução do problema dos mínimos quadrados correspondente ao sistema $Ax = b$, quando a matriz A é mal condicionada ou não tem caracteística máxima. De acordo com o Exercício 3.18, a solução $\widehat{x}$ do problema dos mínimos quadrados é dada por

$$\widehat{x} = \sum_{\sigma_i \neq 0} \frac{u_i^T b}{\sigma_i} v_i,$$

onde os σ_i, u_i e v_i são os valores singulares e os correspondentes vectores singulares de A. Para problemas mal condicionados, os valores singulares pequenos podem ser omitidos na soma e, assim, estabilizar a solução

O problema da compressão de imagens pode, também, ser abordado usando a decomposição em valores singulares de uma matriz. Com efeito, uma imagem a preto e branco pode ser representada por uma matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, onde m e n são, respectivamente, o número de *pixeis* nas direcções horizontal e vertical e o elemento genérico a_{ij} de A representa o nível de cinzento do (i,j)-ésimo *pixel*. Efectuando a decomposição em valores singulares (3.2) de A, obtemos

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_p u_p v_p^T.$$

Podemos aproximar A pela matriz A_k obtida truncando a soma anterior nos k primeiros termos, para $k = 1, 2, \ldots, p$. Como os valores singulares σ_i estão por ordem decrescente, se os últimos $p-k$ forem despresados isso não deverá afectar significativamente a qualidade da imagem. Para transferir a imagem comprimida A_k, bastará transferir os vectores u_i, v_i e os valores singulares σ_i, para $i = 1, 2, \ldots, k$, e não todos os elementos de A.

3.4 Problemas

3.4.1 Exercícios para resolver nas aulas

Exercício 3.2 Seja A uma matriz $n \times n$ da qual se conhece um vector próprio x. Mostre que x é vector próprio associado ao valor próprio $\lambda = \dfrac{\overline{x}^T A x}{\|x\|^2}$, onde $\overline{x}^T$ denota o vector-linha cuja i-ésima componente é igual ao complexo conjugado de x_i.

Exercício 3.3 Verificar que o método da potência não permite calcular o valor próprio de módulo máximo da seguinte matriz, e explicar porquê:

$$A = \begin{bmatrix} 1/3 & 2/3 & 2 & 3 \\ 1 & 0 & -1 & 2 \\ 0 & 0 & -5/3 & -2/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Exercício 3.4 Seja $A = (P^{-1}DP)^T$ com

$$P = \begin{bmatrix} 1 & 1 & 1 \\ 10 & 20 & 30 \\ 100 & 50 & 60 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 60 \end{bmatrix}.$$

Calcule os valores próprios de A.

Exercício 3.5 Usando os círculos de Gershgorin, localize os valores próprios da matriz

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 3 \\ 2 & 4 & 2 \end{bmatrix}$$

e calcule um majorante para o seu raio espectral.

Exercício 3.6 Use os círculos de Gershgorin para provar que os valores próprios de

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

são todos reais.

Exercício 3.7 Considere a matriz

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 4 \\ 2 & 4 & 0 \end{bmatrix}.$$

1. Desenhe os círculos de Gershgorin por linhas e diga o que pode concluir sobre o raio espectral de A (o maior valor próprio de A em valor absoluto).
2. Utilize o método da potência, com vector inicial $x_0 = [1, 0, 0]^T$ para aproximar esse valor (calcule duas iterações do método).

Exercício 3.8 Use o método da potência em segunda aproximação para determinar uma aproximação do valor próprio de módulo máximo da seguinte matriz

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

Exercício 3.9 Localize, usando os círculos de Gershgorin, os valores próprios da seguinte matriz

$$A = \begin{bmatrix} 2 & -1/2 & 0 & -1/2 \\ 0 & 4 & 0 & 2 \\ -1/2 & 0 & 6 & 1/2 \\ 0 & 0 & 1 & 9 \end{bmatrix}.$$

Exercício 3.10 Considere a matriz

$$B = \begin{bmatrix} -5 & 0 & 0{,}5 & 0{,}5 \\ 0{,}5 & 2 & 0{,}5 & 0 \\ 0 & 1 & 0 & 0{,}5 \\ 0 & 0{,}25 & 0{,}5 & 3 \end{bmatrix}.$$

Com base no teorema de Gershgorin obtenha um majorante para o maior módulo de um valor próprio de B e aproxime esse valor próprio efectuando duas iterações do método da potência.

Exercício 3.11 Mostrar que as matrizes $A^{(k)}$ construídas nas iterações do método QR são todas semelhantes à matriz A.

Exercício 3.12 (Matlab) Fixando a tolerância igual a $\varepsilon = 10^{-10}$ e partindo da aproximação inicial $x^{(0)} = [1\ 2\ 3]^T$, usar o método da potência para aproximar o valor próprio de módulo máximo das seguintes matrizes:

$$A_1 = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}; \qquad A_2 = \begin{bmatrix} 0{,}1 & 3{,}8 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}; \qquad A_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Comentar a convergência do método nos três casos.

Exercício 3.13 (Matlab) Verificar que o método da potência não permite calcular o valor próprio de módulo máximo da seguinte matriz, e explicar porquê:

$$A = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 2 & 3 \\ 1 & 0 & -1 & 2 \\ 0 & 0 & -\frac{5}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Exercício 3.14 (Matlab) Têm sido propostos vários modelos matemáticos com o objectivo de prever a evolução de determinadas espécies (humanas ou animais). O modelo de população mais simples, introduzido por Lotka em 1920 e formalizado por Leslie vinte anos mais tarde, é baseado nas taxas de mortalidade e fecundidade para diferentes intervalos de idade, digamos *i*=0,...,n. Seja $x_i^{(t)}$ o número de fêmeas (os machos não intervêm neste contexto) cujas idades no tempo t pertencem ao i-ésimo intervalo. Os valores de $x_i^{(0)}$ são conhecidos. Além disso, seja s_i a taxa de sobrevivência das fêmeas que pertencem ao i-ésimo intervalo, e m_i o número médio de fêmeas geradas por uma fêmea no i-ésimo intervalo de idade.
O modelo de Lotka e Leslie é definido pelas equações

$$x_{i+1}^{(t+1)} = x_i^{(t)} s_i, \quad i = 0, ..., n-1, \qquad x_0^{(t+1)} = \sum_{i=0}^{n} x_i^{(t)} m_i.$$

As n primeiras equações descrevem o desenvolvimento da população, a última a sua reprodução. Em notação matricial, temos

$$x^{(t+1)} = Ax^{(t)},$$

em que $x^{(t)} = \left[x_0^{(t)}, ..., x_n^{(t)}\right]^T$ e A é a matriz de Leslie

$$A = \begin{bmatrix} m_0 & m_1 & ... & ... & m_n \\ s_0 & 0 & ... & ... & 0 \\ 0 & s_1 & \ddots & ... & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & s_{n-1} & 0 \end{bmatrix}.$$

Pode mostrar-se que que a dinâmica desta população é determinada pelo valor próprio de módulo máximo de A, digamos λ_1, enquanto que a distribuição dos indivíduos nos diferentes intervalos de idade (normalizada pela população total), obtém-se como o limite de $x^{(t)}$ para $t \to +\infty$ e verifica $Ax = \lambda_1 x$.

As características de uma população de peixes são descritas pela seguinte matriz de Leslie anteriormente referida:

Intervalo de idade (meses)	$x^{(0)}$	m_i	s_i
$0-3$	6	0	0,2
$3-6$	12	0,5	0,4
$6-9$	8	0,8	0,8
$9-12$	4	0,3	–

.

Determinar o vector x da distribuição normalizada desta população para diferentes intervalos de idade.

Exercício 3.15 (Matlab) Usando os círculos de Gershgorin, dar uma estimativa do número máximo de valores próprios complexos das seguintes matrizes:

$$A = \begin{bmatrix} 2 & -\frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 4 & 0 & 2 \\ -\frac{1}{2} & 0 & 6 & \frac{1}{2} \\ 0 & 0 & 1 & 9 \end{bmatrix}; \qquad B = \begin{bmatrix} -5 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{1}{2} & 3 \end{bmatrix}.$$

Exercício 3.16 (Matlab) Use o comando `eig` para determinar todos os valores próprios das matrizes do exercício anterior.

Exercício 3.17 Considere a matriz de Fibonacci $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$.

1. Determine os valores próprios e os vectores próprios unitários de A^TA e AA^T.
2. Calcule a decomposição em valores singulares da matriz A.

Exercício 3.18 Se $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, com decomposição em valores singulaes $A = U\Sigma V^T$, e $b \in \mathbb{R}$, prove que a solução $\widehat{x}$ do problema dos mínimos quadrados correspondente ao sistema $Ax = b$ é dada por

$$\widehat{x} = V\Sigma^+U^Tb,$$

onde Σ^+ é obtida transpondo Σ e invertendo as suas entradas não nulas.

Exercício 3.19 Prove que a pseudo-inversa A^+ da matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ verifica as chamadas condições de Moore-Penrose:

1. $AA^+A = A$; 2. $A^+AA^+ = A^+$; 3. $(AA^+)^T = AA^+$; 4. $(A^+A)^T = A^+A$.

Exercício 3.20 Mostre que se $A \in \mathcal{M}_n(\mathbb{R})$ é não singular, $A^+ = A^{-1}$.

3.4.2 Exercícios de aplicação à engenharia

Exercício 3.21 Os momentos de inércia, para uma determinada placa, em relação a um sistema coordenado x e y são

$$I_{xx} = 0{,}20 \text{ Kg m}^2, \qquad I_{yy} = 0{,}12 \text{ Kg m}^2,$$

enquanto que os produtos de inércia são

$$I_{xy} = I_{yx} = -0{,}14 \text{ Kg m}^2.$$

O tensor de inércia é representado por

$$J = \begin{bmatrix} I_{xx} & -I_{xy} \\ -I_{yx} & I_{yy} \end{bmatrix}.$$

Os momentos de inércia principal correspondem aos valores próprios de J e os respectivos vectores próprios correspondem aos eixos associados. Use o método da potência para calcular os momentos de inércia principal e a respectiva direcção dos eixos associados.

Exercício 3.22 Considere um corpo formado por três massas: $m_1 = 1$ Kg no ponto $P_1 = \left(\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}\right)$; $m_2 = 2$ Kg no ponto $P_2 = \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, 0, \frac{1}{2}\right)$; $m_3 = 5$ Kg no ponto $P_3 = \left(\frac{1}{2}, -\frac{1}{\sqrt{2}}, \frac{1}{2}\right)$.

1. Calcule o tensor de inércia

$$J = \begin{bmatrix} \sum m_i(y_i^2 + z_i^2) & -\sum m_i x_i y_i & -\sum m_i x_i z_i \\ -\sum m_i y_i x_i) & \sum m_i(x_i^2 + z_i^2) & -\sum m_i y_i z_i \\ -\sum m_i z_i x_i & -\sum m_i z_i y_i & \sum m_i(x_i^2 + y_i^2) \end{bmatrix}.$$

2. Com base no exercício anterior, determine o maior momento de inércia principal e o seu eixo associado, usando o método da potência com erro relativo inferior a 10^{-3} e partindo de $v = [2, –1{,}5, 0]^T$.

Capítulo 4

Equações não lineares

A solução de equações e sistemas de equações é um capítulo em que a análise numérica encontra uma solução bastante precisa. Vamos agora expor alguns métodos que nos permitem obter aproximações para as soluções reais de uma equação real da forma

$$f(x) = 0, \tag{4.1}$$

onde f é uma função que pode ser algébrica ou transcendente.

Os valores de α tais que $f(\alpha) = 0$ são designados por zeros de f, ou raízes de $f(x) = 0$. Só para algumas escolhas particulares de f é que são conhecidos processos que permitem calcular os referidos valores com um número finito de operações.

Exemplo 4.1 As raízes da equação do segundo grau

$$ax^2 + bx + c = 0$$

são facilmente obtidas pela chamada "fórmula resolvente"

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \qquad a \neq 0.$$

Exemplo 4.2 As raízes da equação

$$x^3 + px^2 + qx + r = 0$$

podem ser obtidas pelo processo que se segue, devido a Scipione del Ferro (1465-1515) e Niccolò Tartaglia (1499-1557), publicado pela primeira vez por Gerolamo Cardano (1501-1576). Fazendo a mudança de variável $x = z - \frac{p}{3}$ obtém-se a equação

$$z^3 + az + b = 0,$$

onde

$$a = \frac{1}{3}(3q - p^2) \qquad \text{e} \qquad b = \frac{1}{27}(2p^3 - 9pq + 27r).$$

As raízes desta nova equação são dadas por

$$z_1 = A + B, \qquad z_2 = -\frac{A+B}{2} + \frac{A-B}{2}\sqrt{-3}, \qquad z_3 = -\frac{A+B}{2} - \frac{A-B}{2}\sqrt{-3},$$

onde

$$A = \sqrt[3]{\frac{-b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}, \qquad B = \sqrt[3]{\frac{-b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}.$$

Assim as raízes da equação dada são

$$x_1 = z_1 - \frac{p}{3}; \qquad x_2 = z_2 - \frac{p}{3}; \qquad x_3 = z_3 - \frac{p}{3}.$$

É também possível determinar analiticamente as raízes de uma equação polinomial de quarta ordem. Tal fórmula é devida a Ludovico Ferrari (1522-1569). A fórmula para calcular as raízes de uma equação polinomial de quinta ordem foi procurada durante séculos. Em 1826, o matemático norueguês Niels Henrik Abel (1802-1829) provou que essa fórmula não existe. Assim, para calcular as raízes de uma equação polinomial de grau igual ou superior a cinco temos que recorrer a métodos numéricos. Além disso, de um modo geral, não existem fórmulas para a determinação das raízes de uma equação não polinomial. É o caso que acontece quando consideramos, por exemplo,

$$e^x + \tan x + \log x = 0.$$

A solução analítica de sistemas de equações não lineares também não é possível de obter na maioria dos casos. Como exemplo, considere-se

$$\begin{cases} x^2y + 2xy^2 - xy &= 3 \\ xy^2 - 2x^2y + 4xy &= -1 \end{cases}.$$

Problemas numéricos desta natureza ocorrem com muita frequência na resolução de equações diferenciais, integração, determinação de extremos, etc. Na impossibilidade de obter a sua solução exacta, vamos considerar os chamados métodos iterativos por forma a obter uma solução aproximada para o problema.

4.1 Métodos iterativos

Consideremos o problema (4.1). A filosofia dos métodos iterativos consiste em, partindo de uma aproximação inicial $x^{(0)}$ para uma solução α do problema, gerar uma sucessão de valores

$$x^{(k+1)} = \phi(x^{(k)}), \qquad k = 0, 1, 2, \ldots, \tag{4.2}$$

que seja convergente para essa solução.

Definição 4.1 (Convergência) *O* método iterativo *(4.2) diz-se* convergente *para α se*

$$\lim_{k\to+\infty} |e^{(k)}| = 0,$$

onde $e^{(k)} = e(x^{(k)}) = \alpha - x^{(k)}$ é o erro (absoluto) da iteração k.

Dados vários processos iterativos convergentes para uma solução α de (4.1) coloca- -se a questão de saber qual dos processos é mais eficiente. A eficiência de um processo iterativo pode ser medida de várias maneiras: esforço computacional, tempo gasto, etc. Nesta secção iremos definir um conceito que servirá para medir a velocidade de convergência de um determinado processo iterativo.

Definição 4.2 (Ordem de convergência) *Uma sucessão de iterações $\{x^{(k)}\}$ diz-se que converge com* ordem de convergência *$p \geq 1$ para um ponto α se existir uma constante $M > 0$, independente de k, e uma ordem $k_0 \in \mathbb{N}$ a partir da qual*

$$|e^{(k+1)}| \leq M|e^{(k)}|^p. \tag{4.3}$$

A constante M é chamada constante-erro.

A velocidade de convergência de um processo iterativo está usualmente associada ao conceito de ordem de convergência. Quanto maior for a ordem de convergência mais rápida é, em geral, a velocidade de convergência do processo. A constante-erro também pode ser um aspecto a considerar mas, normalmente, só é tida em conta quando se comparam processos iterativos com a mesma ordem de convergência. Aqui, quanto menor for a constante-erro mais rápida é a convergência do processo.

Se $p = 1$ diz-se que o método iterativo converge linearmente para α. Neste caso a constante erro M terá que ser inferior a 1 (para o método convergir) e a relação (4.3) pode ser escrita na forma $|e^{(k+1)}| \leq M^{k+1}|e^{(0)}|$. Se $p = 2$ diz-se que a convergência é quadrática e se $p = 3$ diz-se que a convergência é cúbica.

Outras questões que surgem naturalmente quando se fala de métodos iterativos são as seguintes: como determinar a aproximação inicial? como definir um método iterativo convergente? como saber que a solução dada pelo método iterativo constitui uma boa aproximação para a solução exacta, isto é, como parar o processo iterativo?

Seja (4.2) o processo iterativo gerador de uma sucessão de aproximações convergente para a solução α de (4.1). Os critérios de paragem mais frequentes, quando se pretende aproximar a raiz α com uma precisão ε, são:

1. Critério do erro absoluto: $|x^{(k)} - x^{(k-1)}| \leq \varepsilon$;
2. Critério do erro relativo: $|x^{(k)} - x^{(k-1)}| \leq \varepsilon |x^{(k)}|$;
3. Critério do valor da função: $|f(x^{(k)})| \leq \varepsilon_1$, onde $\varepsilon_1 \ll \varepsilon$.

Note-se que, se $\{x^{(k)}\}$ for uma sucessão convergente, a sucessão $\{|x^{(k)} - x^{(k-1)}|\}$ também o é e o seu limite é zero. Este facto garante-nos a eficácia dos critérios do erro absoluto e relativo.

Como factor de segurança, para prever o caso em que o processo iterativo possa divergir, também se considera o critério de paragem:

4. Critério do número máximo de iterações: $k = k_{max}$.

4.2 Determinação da aproximação inicial

Num processo iterativo é necessário determinar uma estimativa inicial para a solução do problema a resolver. Por várias razões, algumas delas óbvias, é de todo o interesse que essa aproximação esteja o mais próximo possível da solução exacta. Existem vários processos que permitem encontrar essas aproximações iniciais.

Exemplo 4.3 As soluções de $x^{2,1} - 4x + 2 = 0$ podem ser aproximadas inicialmente pelas soluções de $x^2 - 4x + 2 = 0$.

Exemplo 4.4 Se pretendermos aproximar a maior raiz de $x^5 - x - 500 = 0$ podemos tomar para aproximação inicial $x \approx \sqrt[5]{500} = 3{,}468$.

As técnicas usadas nos exemplos anteriores são muito intuitivas e não podem ser generalizadas a uma gama elevada de problemas. O processo mais usual de obter uma aproximação inicial consiste em tentar obter graficamente um intervalo que contenha a raiz do problema (4.1) que pretendemos calcular. Ora, o traçado gráfico da função f pode não ser evidente e constituir, em si, um processo de complicada resolução sem recurso a *software*

apropriado. Este problema pode ser contornado se reescrevermos a equação (4.1) na forma equivalente

$$f_1(x) = f_2(x), \tag{4.4}$$

sendo f_1 e f_2 funções cujo traçado gráfico seja mais simples que o de f. Assim as raízes de (4.1) serão as soluções de (4.4), isto é, os pontos de intersecção de f_1 com f_2.

O processo de determinação gráfica de um intervalo que contém a raiz deve ser sempre acompanhado de uma confirmação analítica. Para isso, é conveniente relembrar o seguinte teorema devido a Bernard Placidus Johann Nepomuk Bolzano (1781-1848).

Teorema 4.1 (Bolzano) *Se f for uma função contínua em $[a, b]$ então, para todo o y compreendido entre $f(a)$ e $f(b)$, existe pelo menos um $x \in [a, b]$ tal que $f(x) = y$.*

Como pode ser verificado, este teorema estabelece um resultado intuitivo: uma função contínua para passar de um ponto para outro tem de passar por todos os valores intermédios. Como corolário imediato do teorema anterior, temos o seguinte resultado.

Teorema 4.2 (Corolário do Teorema de Bolzano) *Se f for uma função contínua em $[a, b]$ e se $f(a)f(b) < 0$ então existe pelo menos um $c \in]a, b[$ tal que $f(c) = 0$.*

Se, para além das hipóteses do teorema anterior, se verificar que a derivada de f não muda de sinal no intervalo $[a, b]$, então a raiz é única nesse intervalo. Temos assim um critério para verificar a existência e unicidade de zero de uma função contínua f num dado intervalo $[a, b]$: se f é contínua em $[a, b]$, $f(a)f(b) < 0$ e f' não muda de sinal em $[a, b]$, então existe uma e uma só raiz de $f(x) = 0$ em $[a, b]$.

Exercício 4.1 Localize graficamente as raízes de $f(x) = 0$, sendo $f(x) = |x| - e^x$.

Resolução: Como $f(x) = 0 \Leftrightarrow |x| = e^x$, traçando o gráfico de $y = |x|$ e $y = e^x$ (Figura 4.1) verificamos que o seu (único) ponto de intersecção, α (a raiz de $f(x) = 0$), se situa no intervalo $]-1, 0[$.

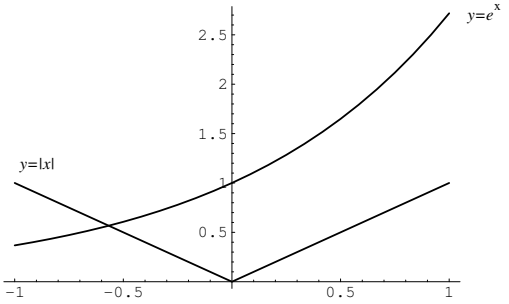

Figura 4.1: Localização gráfica.

De facto, tal acontece uma vez que:

1. $f \in C(]-1, 0[)$;
2. $f(-1)f(0) = 0{,}632 \times (-1) = -0{,}632 < 0$;
3. $f'(x) = -1 - e^x$, para $x < 0$, e como tal $f'(x) < 0$ para todo o $x \in]-1, 0[$.

4.3 Método da bissecção

Seja f uma função contínua em $[a,b]$ tal que $f(a)f(b) < 0$. Então, pelo Teorema 4.2, existe pelo menos uma raiz α de $f(x) = 0$ em $]a,b[$. Se, para além disso, se verificar que a derivada de f não muda de sinal no intervalo $[a,b]$, então a raiz é única nesse intervalo.

Localizada a raiz (localizar uma raiz significa encontrar um intervalo que contenha essa e apenas essa raiz), vamos construir uma sucessão de aproximações convergente para essa raiz. O método mais simples, de entre os que iremos estudar, é o método das divisões sucessivas conhecido por método da bissecção.

No método da bissecção não é necessário o conceito de aproximação inicial mas sim o de intervalo inicial $I^{(0)} =]a,b[=]a^{(0)},b^{(0)}[$. Comecemos por determinar o ponto médio de $I^{(0)}$,

$$x^{(0)} = \frac{a^{(0)} + b^{(0)}}{2}.$$

Caso $f(a^{(0)})f(x^{(0)}) < 0$, temos que $\alpha \in]a^{(0)}, x^{(0)}[$; caso contrário temos que $\alpha \in]x^{(0)}, b^{(0)}[$. Suponhamos, sem perda de generalidade, que $\alpha \in I^{(1)} =]a^{(0)}, x^{(0)}[=]a^{(1)}, b^{(1)}[$. Obtemos assim um intervalo que contém a raiz α de amplitude igual a metade da amplitude do intervalo inicial. Determinando agora o ponto médio de $I^{(1)}$,

$$x^{(1)} = \frac{a^{(1)} + b^{(1)}}{2},$$

podemos obter, de forma análoga, um novo intervalo que contenha a raiz α, de amplitude igual a metade da amplitude do intervalo $I^{(1)}$. Seja esse intervalo $I^{(2)} =]a^{(2)}, b^{(2)}[$. O processo repete-se determinando uma sucessão $\{x^{(k)}\}$ que converge, evidentemente, para α.

O algoritmo do método da bissecção pode ser dado como se segue.

Algoritmo 4.1 Método da bissecção

Dados: a, b, ε_1 e ε_2
Se $f(a)f(b) \geq 0$ então parar
$x := (a+b)/2$
$erro := |b-a|/2$
Enquanto $erro > \varepsilon_1$ e $|f(x)| > \varepsilon_2$ fazer
    Se $f(a)f(x) \leq 0$ então $b := x$ caso contrário $a := x$
    $x := (a+b)/2$
    $erro := |b-a|/2$
Resultado: $\alpha \approx x$

Notemos que, no método da bissecção, a exigência de unicidade de raiz é supérflua. A única exigência é a de que a função tenha sinal contrário nos extremos do intervalo e tal é verificado sempre que exista, nesse intervalo, um número ímpar de raízes.

Verifica-se facilmente que, sendo o intervalo inicial $I^{(0)} =]a,b[$, a amplitude do intervalo $I^{(n)}$ (obtido ao fim de n iterações) é dada por

$$\frac{b-a}{2^n},$$

uma vez que a amplitude do intervalo $I^{(k+1)}$ é sempre igual a metade da amplitude do intervalo $I^{(k)}$, para $k = 1, 2, \ldots$.

Exercício 4.2 Considere o método da bissecção. Seja $[a, b]$ o intervalo que contém uma e uma só raiz α de $f(x) = 0$ e $\{x^{(0)}, x^{(1)}, x^{(2)}, \ldots\}$ a sucessão de pontos médios gerados pelo referido método. Mostre que

1. $|\alpha - x^{(k)}| \leq |x^{(k)} - x^{(k-1)}| = \dfrac{b-a}{2^{k+1}}$.
2. O número, k_{min}, de iterações necessárias para garantir uma aproximação da raiz com uma precisão δ é dado por $k_{min} \geq \log_2 \left(\dfrac{b-a}{\delta} \right) - 1 = -\dfrac{\ln \frac{\delta}{b-a}}{\ln 2} - 1$.

Resolução: 1. Faz-se, sem problemas, por indução.

2. Ao fim de k_{min} iterações obtemos o valor $x^{(k_{min})}$. Assim, pela primeira parte, para calcular qual o k_{min} que verifica $|\alpha - x^{(k_{min})}| \leq \delta$, vamos determinar qual o k_{min} tal que

$$|x^{(k_{min})} - x^{(k_{min}-1)}| = \frac{b-a}{2^{k_{min}+1}} \leq \delta.$$

Temos, sucessivamente,

$$\frac{L}{2^{k_{min}+1}} \leq \delta \Rightarrow \frac{b-a}{\delta} \leq 2^{k_{min}+1} \Rightarrow k_{min} \geq -\frac{\ln \frac{\delta}{b-a}}{\ln 2} - 1.$$

Note-se que, atendendo ao que foi demonstrado no exercício anterior, a convergência do método da bisseção resulta imediatamente uma vez que

$$\lim_{k \to +\infty} |\alpha - x^{(k)}| \leq \lim_{k \to +\infty} \frac{b-a}{2^{k+1}} = 0.$$

Este método possui algumas vantagens bem como algumas desvantagens em relação a outros métodos que iremos estudar nas secções seguintes. A primeira grande vantagem é que o método da bissecção converge sempre (desde que exista raiz no intervalo inicial). A segunda vantagem é que existe uma possibilidade de, *a priori*, se poder indicar um majorante para o erro cometido ao fim de um certo número de iterações.

A grande desvantagem do método da bissecção reside no facto da sua velocidade de convergência ser muito lenta quando comparada com a dos outros métodos. De facto, prova-se que, atendendo à definição de ordem de convergência dada, o método da bissecção converge linearmente e possui uma constante erro $M = \frac{1}{2}$, isto é,

$$|e^{(k+1)}| \leq \frac{1}{2}|e^{(k)}|.$$

Exercício 4.3 É bem sabido que os planetas ao girar em torno do Sol (e os satélites artificiais em torno da Terra) descrevem órbitas elípticas. Para determinar em que ponto da elipse se encontra o móvel num determinado instante t há que resolver a chamada equação de Johannes Kepler (1571-1630)

$$x - e \sin x = z,$$

onde e é a excentricidade (conhecida) da elipse (e que é um valor que varia entre zero, caso a órbita seja circular, e próximo de um, caso a órbita seja muito alongada) e z é um número que se calcula a partir de t. Considerando $e = 0{,}5$ e $z = 0{,}7$, determine a solução do problema com uma casa decimal correcta.

Resolução: Neste caso $f(x) = x - e\sin x - z$, com $e = 0{,}5$ e $z = 0{,}7$. Temos que $f(0) = -0{,}7 < 0$ e $f(2) = 1{,}3 - 0{,}5\sin 2 > 1{,}3 - 0{,}5 = 0{,}8 > 0$. Assim, podemos começar o método da bissecção com o intervalo $I^{(0)} =]0, 2[$. Pelo facto de $f(1) < 0$ temos que a solução pretendida se encontra no intervalo $I^{(1)} =]1, 2[$. Após cinco aplicações do método da bissecção concluímos que a solução se encontra no intervalo $I^{(5)} =]1{,}125, 1{,}1875[$. Tomando como aproximação para a solução o ponto médio $1{,}15625 = (1{,}125+1{,}1875)/2$ temos a garantia que o valor absoluto do erro é inferior a $0{,}03125 < 0.5 \times 10^{-1}$.

Exercício 4.4 Usando o método da bissecção, determine um valor aproximado para o zero de $f(x) = |x| - e^x$, com um erro que não exceda $0{,}15$.

Resolução: Atendendo ao Exercício 4.1, temos que a raiz α de $f(x) = 0$ existe e é única no intervalo $]-1, 0[$. Vamos determinar qual o menor valor de n para o qual $|x^{(n)} - \alpha| \leq 0{,}15$. Pelo Exercício 4.2, esse valor pode ser determinado por

$$\frac{1}{2^{n+1}} \leq 0{,}15 \Leftrightarrow n \geq -\frac{\ln 0{,}15}{\ln 2} - 1 = 1{,}74.$$

Logo, $n = 2$, isto é, temos que efectuar 2 iterações. Partindo do intervalo inicial $]-1, 0[$ temos $x^{(0)} = 0{,}5$. Como $f(x^{(0)}) = -0{,}16065$ vem que

$$\alpha \in]-1, -0{,}5[.$$

Prosseguindo o processo obtemos $x^{(1)} = -0{,}75$ e, como $f(-0{,}75) = 2{,}7776$, vem que

$$\alpha \in]-0{,}75, -0{,}5[.$$

Concluímos então que $\alpha \approx x^{(2)} = -0{,}6256$ é uma aproximação cujo erro não excede a tolerância dada.

4.4 Método de Newton

O método que iremos estudar nesta secção, devido a Newton e a Joseph Raphson ($\sim$1648-$\sim$1715), é um dos métodos mais conhecidos e usados na determinação de aproximações numéricas de raízes de equações não lineares. Para o definir, iremos começar por efectuar uma abordagem analítica fazendo depois a sua interpretação geométrica.

Seja $f \in C^2([a, b])$, com $[a, b] \subset \mathbb{R}$, e $\alpha \in [a, b]$ a única raiz de $f(x) = 0$ nesse intervalo. Pela fórmula de Taylor temos que, se $x^{(0)} \in [a, b]$,

$$f(\alpha) = f(x^{(0)}) + f'(x^{(0)})(\alpha - x^{(0)}) + \frac{f''(\xi)}{2}(\alpha - x^{(0)})^2, \qquad \xi \in I\{\alpha, x^{(0)}\}.$$

Como $f(\alpha) = 0$, e supondo $f'(\alpha) \neq 0$ para todo o $x \in [a, b]$, vem que

$$\alpha = x_0 - \frac{f(x^{(0)})}{f'(x^{(0)})} - \frac{f''(\xi)}{2f'(x^{(0)})}(\alpha - x^{(0)})^2, \qquad \xi \in I\{\alpha, x^{(0)}\}. \tag{4.5}$$

Seja $x^{(1)} = x^{(0)} - \dfrac{f(x^{(0)})}{f'(x^{(0)})}$. Procedendo de forma análoga, poderemos definir um método iterativo pela fórmula de recorrência

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \qquad k = 0, 1 \ldots, \tag{4.6}$$

que pretendemos que seja convergente para α. Este processo iterativo é designado por método de Newton ou método de Newton-Raphson ou ainda método das tangentes. Esta última designação resulta da sua interpretação geométrica.

Interpretação geométrica. Consideremos a recta tangente à curva $y = f(x)$ no ponto de abcissa $x^{(k)}$. Essa recta é dada por

$$y = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}).$$

O ponto de intersecção da recta tangente com o eixo das abcissas é dado por

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

Temos assim que a iteração $x^{(k+1)}$ dada pelo método de Newton é a abcissa do ponto de intersecção da recta tangente à curva $y = f(x)$ no ponto $(x^{(k)}, f(x^{(k)}))$ com a recta $y = 0$.

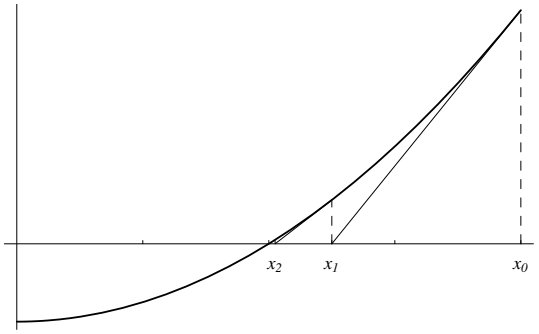

Figura 4.2: Método de Newton.

Vamos agora apresentar as condições que deverão ser impostas para que a sucessão de aproximações geradas pelo método de Newton convirja para a raiz α de $f(x) = 0$.

Teorema 4.3 *Seja f uma função real de variável real definida num intervalo $[a, b]$. Se*

1. $f \in C^2([a, b])$,
2. $f(a)f(b) < 0$,
3. $f'(x) \neq 0$, $x \in [a, b]$,
4. $f''(x) \leq 0$ *ou* $f''(x) \geq 0$, $x \in [a, b]$,

então a sucessão $\{x^{(k)}\}$, $k = 0, 1, \ldots$, gerada pelo método (4.6), com $x^{(0)} \in [a, b]$ tal que

5. $f(x^{(0)})f''(x^{(0)}) > 0$,

converge para a única raiz α de $f(x) = 0$ em $[a, b]$.

Demonstração: Vamos supor, sem perda de generalidade, que $f(a) < 0$, f' é positiva em $[a, b]$ e que f'' é não negativa no mesmo intervalo. Supondo verificadas as hipóteses do teorema, consideremos $x_0 = b$. Provemos que a sucessão $\{x^k\}$ gerada pelo método (4.6) tem as seguintes propriedades.

- A sucessão é não crescente e limitada.

  Vamos provar, por indução, que $x^{(k+1)} \in [\alpha, x^{(k)}]$, para todo o $k \in \mathbb{N}_0$. Por (4.5) tem-se que,

  $$\alpha - x^{(1)} = -\frac{f''(\xi)}{2f'(b)}(\alpha - b)^2 \leq 0, \qquad \xi \in I\{\alpha, b\},$$

isto é, $\alpha \leq x^{(1)}$. Por outro lado, por (4.6), com $k = 0$, tem-se que $x^{(1)} < b$. Suponhamos agora que $x^{(k)} \in [\alpha, x^{(k-1)}] \subseteq [\alpha, b]$. Temos então que, de modo análogo ao efectuado em (4.5),

$$\alpha - \left(x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \right) = -\frac{f''(\xi_k)}{2f'(x^{(k)})}(\alpha - x^{(k)})^2, \qquad \xi_k \in I\{\alpha,^{(k)}\}, \tag{4.7}$$

ou seja

$$\alpha - x^{(k+1)} = -\frac{f''(\xi_k)}{2f'(x^{(k)})}(\alpha - x^{(k)})^2 \leq 0.$$

Isto implica que $\alpha \leq x^{(k+1)}$. Por outro lado, por (4.6) e atendendo às hipóteses do teorema, temos que $x^{(k+1)} - x^{(k)} \leq 0$. Provámos então o pretendido.

- A sucessão converge para α.

  A convergência da sucessão decorre do facto de ela ser não crescente e limitada. Seja $\beta = \lim_{k\to+\infty} x^{(k)}$. Vamos provar que $\beta = \alpha$. Tomando limites em (4.6) e tendo em conta o facto de $f \in C^2([a, b])$, temos que $\beta = \beta - \frac{f(\beta)}{f'(\beta)}$, o que implica $f(\beta) = 0$. Uma vez que α é a única raiz de f em $[a, b]$ temos que $\beta = \lim_{k\to+\infty} x^{(k)} = \alpha$.

Está assim demonstrado o teorema. □

Este método possui vantagens e desvantagens em relação ao método da bissecção. As grandes desvantagens do método de Newton residem no facto deste poder divergir (caso a aproximação inicial escolhida não seja *suficientemente próxima* da raiz) e de haver necessidade de calcular a derivada da função (mais esforço computacional). Por outro lado o método de Newton converge muito rapidamente o que faz com que seja um dos métodos mais eficazes para a aproximação de raízes de equações não lineares.

O teorema seguinte estabelece igualmente uma condição necessária para a convergência do método de Newton. A diferença em relação ao anterior reside apenas na quinta condição: enquanto que o teorema anterior nos dá um critério para a escolha da aproximação inicial, o seguinte dá-nos uma condição que garante a convergência do método para qualquer aproximação inicial escolhida no intervalo $[a, b]$.

Teorema 4.4 *Seja f uma função real de variável real definida no intervalo $[a, b]$. Se*

1. $f \in C^2([a, b])$,
2. $f(a)f(b) < 0$,
3. $f'(x) \neq 0$, $x \in [a, b]$,
4. $f''(x) \leq 0$ *ou* $f''(x) \geq 0$, $x \in [a, b]$,
5. $\left|\frac{f(a)}{f'(a)}\right| \leq b - a$ *e* $\left|\frac{f(b)}{f'(b)}\right| \leq b - a$,

então, qualquer que seja $x^{(0)} \in [a, b]$, a sucessão $\{x^{(k)}\}$ gerada pelo método (4.6) converge para a única raiz α de $f(x) = 0$ em $[a, b]$.

Demonstração: As hipóteses 1, 2 e 3 garantem a existência e unicidade de raiz em $[a, b]$. Provemos que se $x^{(0)} = a$ ou $x^{(0)} = b$ então $x^{(1)} \in]a, b[$. Com efeito, sendo $x^{(0)} = a$ tem-se $x_1 = a - f(a)/f'(a)$ e, da hipótese 5, vem que $-(b-a) < f(a)/f'(a) < b - a$, donde $x^{(1)} < b$. Por outro lado, pelas hipóteses 2 e 3, temos que $f(a)$ tem sinal contrário a $f'(a)$ e como tal $f(a)/f'(a) < 0$. Assim $x^{(1)} - a < 0$ e logo $a < x_1$. De modo idêntico se provaria que se $x^{(0)} = b$ então $x^{(1)} \in]a, b[$.

Suponhamos que $f(a) < 0$. Pela hipótese 4, para $x \in [a, b]$, $f''(x) \leq 0$ ou $f''(x) \geq 0$. Consideremos $f''(x) \leq 0$. Então, de (4.5),

$$\alpha - x^{(1)} = -\frac{f''(\xi)}{2f'(a)}(\alpha - x^{(0)})^2 \geq 0, \qquad \xi \in]a, \alpha[,$$

e, como tal, $x^{(1)} \in]a, \alpha[$. Prova-se também que, nas mesmas condições, $x^{(2)} \in]x^{(1)}, \alpha[$ e, sucessivamente, $x^{(k+1)} \in]x^{(k)}, \alpha[$, $k = 0, 1, \ldots$.

Provámos assim que a sucessão $\{x^{(k)}\}$ converge monotonamente para α.

Os restantes casos podem ser considerados de forma análoga. □

O algoritmo para o método de Newton pode ser dado como se segue.

Algoritmo 4.2 Método de Newton

Dados: $x^{(0)}$, ε e k_{max}

$x := x^{(0)}$

$k := 0$

$erro := \varepsilon + 1$

Enquanto $erro > \varepsilon$ e $k < k_{max}$ fazer

$k := k + 1$

$d := -f(x)/f'(x)$

$x := x + d$

$erro := |d|$

Resultado: $\alpha \approx x$

Não é difícil provar a convergência quadrática do método de Newton. De facto, tomando módulos em (4.7) obtemos

$$|e^{(k+1)}| \leq M|e^{(k)}|^2,$$

com

$$M = \frac{1}{2}\frac{\max_{x\in[a,b]} |f''(x)|}{\min_{x\in[a,b]} |f'(x)|}. \tag{4.8}$$

Assim, supondo verificadas as hipóteses do Teorema 4.3, concluímos que o método de Newton tem ordem de convergência $p = 2$.

Outra vantagem do método de Newton em relação ao método da bissecção tem a ver com o facto do método de Newton se poder generalizar muito facilmente (como veremos) para sistemas de equações não lineares. Além disso, este método também se pode aplicar ao cálculo numérico de raízes complexas.

Exercício 4.5 Localize graficamente as raízes de $f(x) = 0$, onde

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

e aproxime a maior delas usando o método de Newton duas vezes.

Resolução: Como $f(x) = 0 \Leftrightarrow x^2 - 1 = \ln(x+1)$, traçando o gráfico de $y = x^2 - 1$ e de $y = \ln(x+1)$ (Figura 4.3) verificamos que $f(x) = 0$ possui duas raízes reais: $\alpha_1 \in]-1, 0[$ e $\alpha_2 \in]1, 2[$.

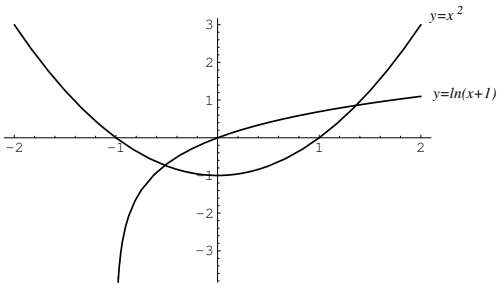

Figura 4.3: Localização gráfica.

Façamos a confirmação analítica apenas para α_2. Assim:

1. $f \in C(]1, 2[)$;
2. $f(1) = -\ln 2 < 0$ e $f(2) = 3 - \ln 3 = 1{,}901388 > 0$;
3. $f'(x) = 2x - (x-1)^{-1} > 0$, para $x \in]1, 2[$.

Logo a raiz α_2 de $f(x) = 0$ existe e é única no intervalo $[1, 2]$.

Para aplicarmos o método de Newton temos primeiro que provar a sua convergência. Como $f(x) = x^2 - 1 - \ln(x+1)$, $f'(x) = 2x - (x+1)^{-1}$ e $f''(x) = 2 + (x+1)^{-2}$ temos que $f \in C^2([1, 2])$. Por outro lado, como $f'(x) > 0$ (prove!) e $f''(x) \geq 0$, para todo o $x \in [1, 2]$, o Teorema 4.3 garante que o método de Newton aplicado à equação dada gera uma sucessão de valores convergentes para α_2, desde que $x^{(0)}$ seja escolhido por forma a que $f(x^{(0)})f''(x^{(0)}) > 0$, isto é, por forma a que $f(x^{(0)}) > 0$.

Seja, então, $x^{(0)} = 2$. Assim

$$x^{(1)} = 2 - \frac{f(2)}{f'(2)} = 1{,}48144;$$

$$x^{(2)} = 1{,}48144 - \frac{f(1{,}48144)}{f'(1{,}48144)} = 1{,}369785.$$

Uma estimativa para o erro absoluto pode ser dada por $|x^{(2)} - x^{(1)}| = 0{,}1116554$.

4.5 Método do ponto fixo

O método do ponto fixo não é propriamente um método mas sim uma classe de métodos (o método de Newton, por exemplo, pertence à classe de métodos do ponto fixo). Este método tem grande importância na resolução de todo o tipo de equações, incluindo as equações diferenciais e integrais. Neste momento vamos apenas considerar o problema da determinação das raízes de uma equação não linear $f(x) = 0$.

O método do ponto fixo consiste em converter o problema de determinar os zeros de uma função no problema (equivalente) de calcular os pontos fixos de uma outra função.

Definição 4.3 (Ponto Fixo) *Seja ϕ uma função definida num intervalo real $[a,b]$. Dizemos que $\alpha \in [a,b]$ é um* ponto fixo *de ϕ se $\alpha = \phi(\alpha)$.*

Assim, o problema de determinar os valores de x para os quais $f(x) = 0$ (zeros de f) é transformado no problema equivalente de determinar os valores de x para os quais $\phi(x) = x$ (pontos fixos de ϕ). Consideremos o seguinte exemplo.

Exemplo 4.5 A excentricidade da órbita de Vénus é dada por $e = 0{,}07$. Suponhamos que pretendemos resolver a equação de Kepler $x - 0{,}007 \sin x - z = 0$, quando $z = 0{,}7$. Como o termo $0{,}007 \sin x$ é muito menor que $0{,}7$ temos que uma aproximação para a solução pode ser dada por $x \approx 0{,}7$. Substituindo este valor em $0{,}007 \sin x$ obtemos $0{,}007 \sin 0.7 \approx 0{,}004510$. Introduzindo este valor na equação de Kepler temos uma nova aproximação para a sua raiz dada por $x \approx 0{,}7 + 0{,}004510 = 0{,}704510$. Este processo poderia continuar dando assim origem a um processo iterativo da forma $x^{(k+1)} = \phi(x^{(k)})$, $k = 0, 1, \ldots$, com $\phi(x) = 0{,}7 + 0{,}007 \sin x$ e $x^{(0)} = 0{,}7$.

Depois de transformar o problema na forma da determinação dos pontos fixos de uma função ϕ, as sucessivas aproximações são calculadas, a partir de uma aproximação inicial $x^{(0)}$ dada, pela fórmula

$$x^{(k+1)} = \phi(x^{(k)}), \qquad k = 0, 1, 2, \ldots. \tag{4.9}$$

A função ϕ é chamada função de iteração do método. Notemos que, no caso do método de Newton, a função de iteração é dada por

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

A questão que se coloca é a seguinte: dada uma equação $f(x) = 0$ com raiz $\alpha \in [a,b]$, como escolher uma função de iteração ϕ por forma a que as sucessivas aproximações dadas por (4.9) convirjam para α? Antes de mais notemos que, supondo que ϕ é contínua e que $x^{(k)} \to \alpha$, se tem

$$\alpha = \lim_{k \to +\infty} x^{(k+1)} = \lim_{k \to +\infty} \phi(x^{(k)}) = \phi\left(\lim_{k \to +\infty} x^{(k)}\right) = \phi(\alpha).$$

Assim, uma condição necessária para que o processo iterativo (4.9) convirja para zero α de f é que α seja um ponto fixo de ϕ.

Exemplo 4.6 A equação de Kepler dada no exemplo anterior pode escrever-se na forma

$$x = \phi(x) = \arcsin\left(\frac{x - 0{,}7}{0{,}007}\right).$$

Neste caso, para a aproximação inicial $x^{(0)} = 0{,}7$ temos que $x^{(1)} = 0$ e $x^{(2)} = \arcsin(-100)$ que é um valor que nem sequer está definido. Como tal, esta escolha para a função de iteração não é adequada.

Como poderemos decidir qual a melhor escolha para a função de iteração? Em geral, interessa que $\phi(x)$ varie pouco com x. O caso ideal seria ter ϕ constante; nesse caso, para $x^{(0)}$ arbitrário, teríamos $x^{(1)} = \alpha$. Para responder a esta questão, consideremos o seguinte teorema.

Teorema 4.5 (Ponto Fixo) *Seja ϕ uma função real de variável real definida no intervalo $[a,b]$. Se*

1. *ϕ é uma função contínua em $[a,b]$ e*
2. *$\phi(x) \in [a,b]$ para todo o $x \in [a,b]$,*

então ϕ tem um ponto fixo $\alpha \in [a,b]$. Se, além disso, ϕ é diferenciável em $]a,b[$ e

3. *$|\phi'(x)| \leq K < 1$, para todo o $x \in]a,b[$,*

então o ponto fixo é único e a sucessão gerada por (4.9) converge para esse ponto, qualquer que seja a aproximação inicial $x^{(0)} \in [a,b]$. Além disso

$$\lim_{k\to+\infty} \frac{x^{(k+1)}-\alpha}{x^{(k)}-\alpha} = \phi'(\alpha). \tag{4.10}$$

Demonstração: Vamos mostrar sucessivamente a existência de ponto fixo, a unicidade e, finalmente, a convergência do método.

- Existência.

  Se $\phi(a) = a$ ou $\phi(b) = b$ temos que ϕ tem (obviamente) um ponto fixo. Caso contrário, e atendendo à hipótese 2 do teorema, temos que $\phi(a) > a$ e $\phi(b) < b$. Consideremos a função auxiliar $\psi(x) = \phi(x) - x$ definida em $[a,b]$. Como ψ é contínua e $\psi(a)\psi(b) < 0$ concluímos que existe um ponto $\alpha \in [a,b]$ tal que $\psi(\alpha) = 0$, ou seja, tal que $\phi(\alpha) = \alpha$.

- Unicidade.

  Suponhamos que α_1 e α_2 são dois pontos fixos de ϕ. Então

  $$|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| = |\phi'(\eta)||\alpha_1 - \alpha_2| \leq K|\alpha_1 - \alpha_2|,$$

  onde η pertence ao intervalo definido por α_1 e α_2. Assim sendo $(1-K)|\alpha_1 - \alpha_2| \leq 0$, o que implica $\alpha_1 = \alpha_2$, uma vez que $0 \leq K < 1$.

- Convergência e (4.10).

  Considerando $x^{(0)} \in [a,b]$ temos que

  $$\begin{aligned} |x^{(k+1)} - \alpha| &= |\phi(x^{(k)}) - \phi(\alpha)| \\ &= |\phi'(\xi_k)||x^{(k)} - \alpha| \leq K|x^{(k)} - \alpha|, \quad \xi_k \in I\{\alpha, x^{(k)}\}. \end{aligned} \tag{4.11}$$

  Assim sendo

  $$|x^{(k+1)} - \alpha| \leq K^{k+1}|x^{(0)} - \alpha|. \tag{4.12}$$

  Tomando limites e atendendo a que $K < 1$ temos que

  $$\lim_{k\to+\infty} x^{(k+1)} = \alpha,$$

  o que prova a convergência do método. Para provar (4.10), notemos que, de (4.11) sai que

  $$\frac{x^{(k+1)}-\alpha}{x^{(k)}-\alpha} = \phi'(\xi_k), \qquad \xi_k \in I\{\alpha, x^{(k)}\}.$$

  Como o método converge para α, tomando limites em ambos os membros, provamos o pretendido. □

Notemos que, atendendo a (4.10), o método do ponto fixo tem, no caso geral, uma convergência linear. Além disso, essa convergência é local, uma vez que ela só acontece quando o $x^{(0)}$ está suficientemente próximo do ponto fixo.

Exemplo 4.7 Resolvamos, mais uma vez, a equação de Kepler considerando: (i) a excentricidade $e = 0{,}5$ e $z = 0{,}7$; (ii) a excentricidade $e = 0{,}5$ e $z = 2$. Vamos apenas efectuar os cálculos para o caso (ii), isto é, vamos considerar apenas a equação $x - 0{,}5 \sin x - 2 = 0$. Para usar o método do ponto fixo consideremos a função de iteração

$$\phi(x) = 0{,}5 \sin x + 2, \qquad x \in [2, 3].$$

Vejamos se, para esta função e para este intervalo, se verificam as condições de convergência do método. Como ϕ é uma função contínua, vamos provar que $\phi(x) \in [2, 3]$, para todo o $x \in [2, 3]$, isto é, que o gráfico de ϕ está totalmente contido no quadrado $[2, 3] \times [2, 3]$. Para isso temos que provar que $\phi(2), \phi(3) \in [2, 3]$ e que o valor ϕ em todos os seus extremos locais também se encontra nesse intervalo. Ora, $\phi(2) = 2{,}4546$, $\phi(3) = 2{,}0706$ e a função ϕ é monótona decrescente (pois $\phi'(x) = 0{,}5 \cos x$). Assim sendo, $\phi(x) \in [2, 3]$, para todo o x a variar nesse intervalo. Para provar que o método converge basta apenas provar que o majorante do módulo de ϕ', em $]2, 3[$, é inferior a um. Como se vê facilmente $|\phi'(x)| = |0{,}5 \cos x| \leq 0{,}5$, e, como tal, $K = 0{,}5$ e o método $x^{(k+1)} = 0{,}5 \sin x^{(k)} + 2$, $k = 0, 1, 2, \ldots$, converge para a única raiz da equação em $[2, 3]$, qualquer que seja $x^{(0)} \in [2, 3]$. A determinação das sucessivas iterações é feita de forma óbvia.

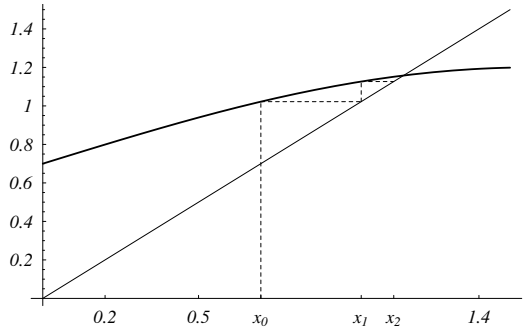

Figura 4.4: Caso (i): $\phi(x) = 0{,}5 \sin x + 0{,}7$ e $x^{(0)} = 0{,}7$.

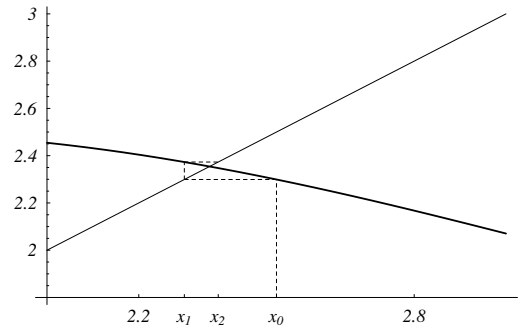

Figura 4.5: Caso (ii): $\phi(x) = 0{,}5 \sin x + 2$ e $x^{(0)} = 2{,}5$.

Nas Figuras 4.4 e 4.5 podemos visualizar o gráfico da função de iteração do método do ponto fixo considerado no exemplo anterior, para os casos (i) e (ii), respectivamente, bem como o gráfico da bissectriz dos quadrantes ímpares. A abcissa da intersecção dos dois gráficos é o ponto fixo que pretendemos calcular. Como se pode ver, o método do ponto fixo pode ser descrito da seguinte forma (gráfica). A partir de $x^{(0)}$, no eixo das abcissas, traçamos um segmento de recta vertical até intersectar o gráfico de ϕ. A ordenada da intersecção é o ponto $x^{(1)}$. A partir do ponto de intersecção traçamos um segmento de recta horizontal até encontrarmos a bissectriz $y = x$. A abcissa desse ponto final é $x^{(1)}$. Para determinar as restantes iterações repete-se sucessivamente este processo: vertical até ao gráfico de ϕ, horizontal até à bissectriz.

Regressemos, de novo, à questão de saber qual a melhor escolha para a função de iteração. O Teorema do Ponto Fixo permite-nos afirmar que se uma função de iteração não verificar as hipóteses do teorema, essa função não deve ser considerada. Pode, no entanto, dar-se o caso de possuirmos duas funções de iteração que verifiquem, ambas, as hipóteses do teorema. Neste caso, por qual optar? Notemos que, por (4.12), se considerarmos duas funções de iteração ϕ_1 e ϕ_2 tais que

$$|\phi_1'(x)| \leq |\phi_2'(x)| < 1, \qquad x \in]a, b[,$$

podemos concluir que a sucessão definida pelo método $x^{(k+1)} = \phi_1(x^{(k)})$, $k = 0, 1, \ldots$, converge mais rapidamente do que a sucessão definida por $x^{(k+1)} = \phi_2(x^{(k)})$, $k = 0, 1, \ldots$, pois para o primeiro método temos $|e^{(k+1)}| \leq M_1|e^{(k)}|$ e para o segundo $|e^{(k+1)}| \leq M_2|e^{(k)}|$, com $M_1 \leq M_2$. Assim sendo, a escolha deveria recair sobre a função ϕ_1.

Como vimos, o método do ponto fixo tem convergência linear. No entanto, o método de Newton (caso particular do método do ponto fixo quando a função de iteração é dada por $\phi_N(x) = x - f(x)/f'(x)$) tem convergência quadrática. O próximo teorema diz-nos em que condições podemos garantir uma ordem de convergência dois no método do ponto fixo.

Teorema 4.6 *Suponhamos que, para além das hipóteses do Teorema do Ponto Fixo, se tem $\phi'(\alpha) = 0$ (onde α é o único ponto fixo de ϕ em $[a, b]$), e ϕ'' limitada em $]a, b[$. Então o método do ponto fixo (4.9) converge para α de forma quadrática, qualquer que seja $x^{(0)} \in [a, b]$. Além disso*

$$\lim_{k \to +\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \phi''(\alpha). \tag{4.13}$$

Demonstração: Pelo Teorema do Ponto Fixo temos que o método (4.9) converge para α. Falta apenas provar que a convergência é quadrática.

Pela fórmula de Taylor temos que

$$x^{(k+1)} - \alpha = \phi'(\alpha)(x^{(k)} - \alpha) + \frac{1}{2}\phi''(\xi_k)(x^{(k)} - \alpha)^2, \qquad \xi_k \in I\{\alpha, x^{(k)}\}.$$

Como $\phi'(\alpha) = 0$ concluímos que $|e^{(k+1)}| \leq M|e^{(k)}|^2$, onde $M = \frac{1}{2}\max_{x \in [a,b]} |\phi''(x)|$. Está, assim, demonstrado que o método (4.9) tem ordem 2. A demonstração de (4.13) é semelhante à efectuada para demonstrar (4.10). □

Exercício 4.6 Mostre que, se no ponto fixo α de ϕ se tem $\phi'(\alpha) = \phi''(\alpha) = 0$, podemos concluir (mediante certas condições) que o método (4.9) tem convergência cúbica. Diga quais são essas condições de convergência.

Consideremos agora os seguintes corolários do Teorema do Ponto Fixo, úteis para determinar estimativas *a priori* para o erro cometido ao fim de um determinado número de iterações.

Corolário 4.7 *Nas hipóteses do Teorema do Ponto Fixo tem-se que*

$$|e^{(k)}| \leq K^k \max\{x^{(0)} - a, b - x^{(0)}\}.$$

Demonstração: Resulta imediatamente de (4.12). □

Corolário 4.8 *Nas hipóteses do Teorema do Ponto Fixo tem-se que*

$$|e^{(k)}| \leq \frac{K^k}{1-K}|x^{(1)} - x^{(0)}|.$$

Demonstração: Por um processo análogo ao efectuado na demonstração do Teorema do Ponto Fixo temos que

$$|x^{(k+1)} - x^{(k)}| \leq K^k|x^{(1)} - x^{(0)}|.$$

Consideremos $l > k$ e $|x^{(l)} - x^{(k)}|$. Assim

$$|x^{(l)} - x^{(k)}| \leq \sum_{j=k}^{l-1} |x^{(j+1)} - x^{(j)}| \leq |x^{(1)} - x^{(0)}| \sum_{j=k}^{l} K^j.$$

Logo

$$\sum_{j=k}^{l} K^j \leq K^k \sum_{j=0}^{\infty} K^j = \frac{K^k}{1-K}.$$

Concluímos então que

$$|x^{(l)} - x^{(k)}| \leq \frac{K^k}{1-K}|x^{(1)} - x^{(0)}|.$$

Tomando o limite quando $l \to +\infty$ temos

$$|\alpha - x^{(k)}| \leq \frac{K^k}{1-K}|x^{(1)} - x^{(0)}|,$$

o que prova o pretendido. □

4.6 Equações algébricas

Suponhamos agora que pretendemos resolver a equação algébrica $P_n(x) = 0$ onde

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \qquad a_n \neq 0, \tag{4.14}$$

é um polinómio de coeficientes reais. Este problema aparece com muita frequência e existem para ele muitos resultados específicos. Nesta secção faremos apenas uma breve referência a alguns desses resultados.

Um resultado básico sobre polinómios é dado no Teorema Fundamental da Álgebra, devido a Gauss e a Euler e que apresentamos sem demonstração.

Teorema 4.9 (Teorema Fundamental da Álgebra) *Seja P_n um polinómio de grau n, com $n \geq 1$, de coeficientes reais. Então existe $\alpha \in \mathbb{C}$ tal que $P_n(\alpha) = 0$.*

Temos também que, no caso particular dos polinómios, se α é um zero real de P_n então

$$P_n(x) = (x - \alpha)Q_{n-1}(x; \alpha),$$

onde Q_{n-1} é um polinómio de grau $n-1$ de coeficientes reais que dependem de α. Se α é um zero complexo de P_n o seu conjugado $\overline{\alpha}$ também o é e, como tal,

$$P_n(x) = (x - \alpha)(x - \overline{\alpha})Q_{n-2}(x; \alpha),$$

sendo Q_{n-2} um polinómio de grau $n-2$ de coeficientes reais. Atendendo a estes resultados podemos escrever.

Corolário 4.10 *Um polinómio de grau $n \geq 1$ de coeficientes reais admite, exactamente, n zeros, reais ou complexos, iguais ou distintos.*

Corolário 4.11 *Se P_n for um polinómio de grau ímpar admite, pelo menos, uma raiz real.*

A localização das raízes reais de uma equação algébrica pode ser feita por variadíssimos processos. De entre os processos mais populares destaca-se o método de Rolle. A justificação teórica do método é dada por um corolário do seguinte teorema estabelecido por Michel Rolle (1652-1719).

Teorema 4.12 (Rolle) *Se f for uma função contínua em $[a, b]$, diferenciável em $]a, b[$ e se $f(a) = f(b)$ então existe pelo menos um $\xi \in]a, b[$ tal que $f'(\xi) = 0$.*

Notemos que, quando $f(a) = f(b) = 0$, este teorema diz-nos, em linguagem comum, que entre dois zeros de uma função contínua existe, pelo menos, um zero da sua derivada.

O corolário que importa considerar neste contexto, é o seguinte.

Teorema 4.13 (Corolário do Teorema de Rolle) *Se f for uma função contínua num intervalo $[a, b]$ e diferenciável em $]a, b[$ e se a e b são dois zeros consecutivos de f', então existe, no máximo, um $\xi \in]a, b[$ tal que $f(\xi) = 0$.*

Este teorema, em linguagem (muito) informal, costuma ser enunciado de forma seguinte: entre dois zeros consecutivos da derivada de uma dada função, existe, no máximo, um zero dessa função.

Para definir o método de Rolle consideremos, previamente, a seguinte definição.

Definição 4.4 (Números de Rolle) *Chamam-se* números de Rolle *da equação $f(x) = 0$, definida em $I \subseteq \mathbb{R}$, ao conjunto formado pelos pontos fronteira de I e pelos zeros da derivada de f.*

Atendendo ao teorema anterior temos que, uma vez ordenados de forma crescente, entre dois números de Rolle consecutivos existe, no máximo, uma raiz real da equação. Assim se o valor da função tiver o mesmo sinal nos extremos do intervalo definido por dois números de Rolle consecutivos, a equação não tem nenhuma raiz real nesse intervalo; caso contrário, a equação tem uma só raiz real no intervalo.

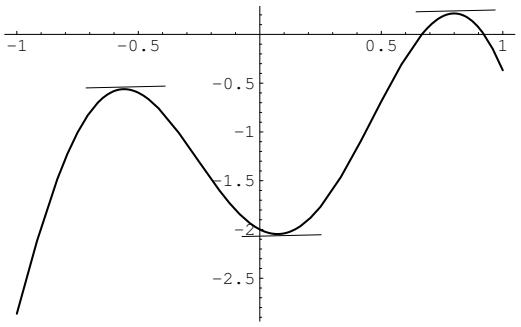

Figura 4.6: Corolário do Teorema de Rolle.

Exercício 4.7 Usando o método de Rolle, localize todas as raízes reais de

$$P_3(x) \equiv x^3 - 2x - 5 = 0.$$

Resolução: A função derivada, definida pela expressão $3x^2 - 2$, tem dois zeros $-\sqrt{2/3}$ e $+\sqrt{2/3}$. Os números de Rolle da equação dada são

$$-\infty, \qquad -\sqrt{\frac{2}{3}}, \qquad +\sqrt{\frac{2}{3}}, \qquad +\infty.$$

Como a função dada é contínua em $\mathbb{R}$ e

r_i	$-\infty$		$-\sqrt{2/3}$		$+\sqrt{2/3}$		$+\infty$
$P_3(r_i)$		$-$		$-$		$+$	

,

temos que a única raiz real da equação dada está no intervalo $]\sqrt{2/3}, +\infty[$.

Note-se que o intervalo que contém todas as raízes da equação algébrica pode ser obtido recorrendo ao seguinte resultado.

Teorema 4.14 (Cauchy) *Seja $P_n(x) = 0$ uma equação algébrica, com P_n um polinómio da forma (4.14). Todos os zeros de P_n estão incluídos no círculo Γ do plano complexo*

$$\Gamma = \{z \in \mathbb{C} : |z| \leq 1 + \eta\}, \quad com \quad \eta = \max_{k=0,\ldots,n-1} \left|\frac{a_k}{a_n}\right|.$$

Um resultado alternativo é o seguinte.

Teorema 4.15 (Newton) *Seja $P_n(x) = 0$ uma equação algébrica. Se, para $x = L$, $L > 0$, o polinómio P_n e as suas sucessivas derivadas forem não negativas, então L constitui um limite superior das raízes positivas de $P_n(x) = 0$.*

Demonstração: Seja P_n um poliómio de grau n. Fazendo o seu desenvolvimento de em série Taylor, em torno de $x = L$, temos que

$$P_n(x) = P_n(L) + P_n'(L)(x - L) + \frac{P_n''(L)}{2!}(x - L)^2 + \cdots + \frac{P_n^{(n)}(L)}{n!}(x - L)^n.$$

Assim é fácil concluir que, nas hipótese do teorema, $P_n(x) > 0$ para todo o $x > L$, o que prova o pretendido. □

Um limite inferior l para as raízes negativas de $P_n(x) = 0$ poderia ser obtido usando o resultado anterior, atendendo a que as raízes negativas de uma equação algébrica $P_n(x) = 0$, onde P_n é um polinómio de grau n, são as raízes positivas, com sinal contrário, de $Q_n(x) \equiv (-1)^n P_n(-x) = 0$.

Exercício 4.8 Prove a afirmação anterior.

Exercício 4.9 Determine limites superiores e inferiores para as raízes reais de $x^3 - 2x - 5 = 0$.

Resolução: Seja $P_3(x) = x^3 - 2x - 5$. Atendendo a que

	0	1	2	3
$P_3(x)$	$-$	$-$	$-$	$+$
$P_3'(x)$				$+$
$P_3''(x)$				$+$
$P_3'''(x)$				$+$

,

$L = 3$ é limite superior das raízes de $P_3(x) = 0$. Para determinar um limite inferior das raízes, consideremos $Q_3(x) \equiv (-1)^3 P_3(-x) = x^3 - 2x + 5$. Ora, atendendo a que

	0	1
$Q_3(x)$	$-$	$+$
$Q_3'(x)$		$+$
$Q_3''(x)$		$+$
$Q_3'''(x)$		$+$

,

temos que $l = -1$ é limite inferior das raízes de $P_3(x) = 0$.

Outro resultado muito útil para determinar o número de zeros reais positivos de um polinómio foi enunciado por René Descartes (1596-1650) em 1637: "O número de zeros reais positivos de um polinómio é limitado pelo número de varições de sinal da sucessão dos seus coeficientes". Mais tarde, Gauss demonstrou que "o número de zeros reais positivos de um poliómio (contando com a multiplicidade) tem a mesma paridade do número de variações de sinal da sucessão dos seus coeficientes". Temos então o seguinte teorema.

Teorema 4.16 (Regra de Sinal de Descartes) *O número de raízes reais positivas da equação $P_n(x) = 0$, sendo P_n dado por (4.14), é igual ao número de variações de sinal da sucessão $\{a_n, a_{n-1}, \ldots, a_0\}$ ou um número inferior mas da mesma paridade.*

Demonstração: Vamos efectuar a demonstração por indução.

Comecemos por considerar $n = 1$, isto é, P_n um polinómio de grau um. Neste caso o resultado é óbvio pois a raiz de $P_n(x) = 0$, com $P_n(x) = a_1 x + a_0$, só é positiva quando e só quando $a_1 a_0 > 0$.

Suponhamos agora que o resultado é válido para todos os polinómios de grau $n - 1$ e consideremos P_n um polinómio de grau n dado por (4.14), com $a_n > 0$ (sem perda de generalidade). Se $a_0 = P_n(0) > 0$, o número de variações de sinal da sucessão dos coeficientes de P_n tem que ser par pois o primeiro e o último termo da sucessão são positivos. Por outro lado, o número de raízes positivas de $p_n(x) = 0$ também é par pois $\lim_{x \to +\infty} P_n(x) = +\infty$.

A mesma argumentação poderia ser usada no caso de $a_0 = P_n(0) < 0$; neste caso, tanto o número de varições de sinal da sucessão dos coeficientes de P_n como o número de

zeros positivos de P_n são ímpares. Concluímos então que o número de raízes positivas de $P_n(x) = 0$ tem a mesma paridade do número de variações de sinal.

Falta apenas provar que o número de variações de sinal limita o número de raízes positivas. Suponhamos que $P_n(x) = 0$ tem m raízes reais positivas e que o número de variações de sinal da sucessão dos seus coeficientes é $V < m$. Assim sendo, temos que ter $m \geq V + 2$ (para manter a paridade). Mas, pelo Teorema de Rolle, P_n' tem que ter pelo menos $V + 1$ raízes reais positivas, o que contraria a hipótese de indução uma vez que o número de variações de sinal dos coeficientes de P_n' (polinómio de grau $\leq n - 1$) é inferior a V. Logo $m \leq V$. □

Notemos que a regra de sinal de Descartes não tem em conta a multiplicidade das raízes. No entanto, podemos afirmar o resultado demonstrado por Gauss, isto é que o número de raízes reais positivas de $P_n(x) = 0$ (contando com a multiplicidade) tem a mesma paridade do número de variações de sinal da sucessão dos seus coeficientes.

Exercício 4.10 Usando a regra de sinal de Descartes, determine o número de raízes reais de $P_3(x) = 0$, onde $P_3(x) = x^3 - 2x - 5$.

Resolução: Comecemos pelas raízes positivas. Como a sucessão de sinais dada pelos coeficientes do polinómio é $\{+, -, -\}$, temos que o número de variações de sinal é 1 e, como tal, existe uma raiz positiva de $P_3(x) = 0$. Para as raízes negativas consideremos o polinómio auxiliar $Q_3(x) = (-1)^3 P_3(-x) = x^3 - 2x + 5$. Como a sucessão de sinais dada pelos coeficientes do polinómio Q_3 é $\{+, -, +\}$, temos que o número de variações de sinal é 2 e, como tal, existem 2 ou 0 raízes negativas de $P_3(x) = 0$.

4.6.1 Algoritmo de Hörner

O cálculo dos zeros de um polinómio é feito, na maioria das vezes, recorrendo ao método de Newton. Quando se aplica este método há necessidade de calcular, em cada iteração, o valor do polinómio e da sua derivada num ponto. Esse cálculo deve ser efectuado de forma eficiente uma vez que grande parte do esforço computacional a ele se deve.

Suponhamos que se pretende calcular $P_n(z)$, com P_n um polinómio dado por (4.14). Considerando o polinómio escrito na forma canónica (tal como em (4.14)), efectuamos n adições/subtracções e $2n - 1$ multiplicações/divisões. Considerando a forma encaixada

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + a_n x))),$$

ao calcular $P_n(z)$ só efectuamos n adições/subtracções e n multiplicações/divisões. Esta forma está na base do algoritmo de William George Hörner (1786-1837).

Algoritmo 4.3 Algoritmo de Hörner

Dados: a_i, $i = 0, 1, \ldots, n$, e z

$b := a_n$

Para k de $n - 1$ até 0 fazer

    $b := a_k + bz$

Resultado: $P_n(z) = b$

Se P_n for um polinómio dado por (4.14) e z um número real temos que

$$P_n(x) = (x - z)Q_{n-1}(x; z) + b_0, \tag{4.15}$$

onde Q_{n-1} é um polinómio de grau $n - 1$, que depende de z, dado por

$$Q_{n-1}(x; z) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1, \tag{4.16}$$

designado por polinómio associado a P_n. Note-se que o valor de $P_n(z) = b_0$, ou seja, temos o seguinte resultado demonstrado por Paolo Ruffini (1765-1822).

Teorema 4.17 (Ruffini) *O valor numérico de $P_n(z)$ de um polinómio P_n em $x = z$ é igual ao resto da divisão de $P_n(x)$ por $(x - z)$.*

A chamada regra de Ruffini, que consiste em substituir (4.14) e (4.16) em (4.15) e igualando os coeficientes de potências de x do mesmo grau, permite obter os valores b_i, $i = 1, \ldots, n$, e o valor de b_0 de acordo com o algoritmo de Hörner. O mesmo algoritmo permite obter facilmente os valores da derivada de P_n dado por (4.14) num dado ponto z. Assim, uma vez que, por (4.15),

$$P'_n(x) = (x - z)Q'_{n-1}(x; z) + Q_{n-1}(x; z),$$

temos que $P'_n(z) = Q_{n-1}(z; z)$.

Algoritmo 4.4 Valores da derivada de um polinómio

Dados: a_i, $i = 0, 1, \ldots, n$, e z
$q := 0$
$b := a_n$
Para k de $n - 1$ até 0 fazer
    $q := b + qz$
    $b := a_k + bz$
Resultado: $P_n(z) = b$ e $P'_n(z) = q$

4.6.2 O método de Newton-Hörner

Vamos começar por considerar o caso em que P_n, dado por (4.14), tem apenas zeros reais simples. Neste caso, podemos aplicar qualquer um dos métodos iterativos estudados. No entanto, sugerimos o seguinte procedimento.

1. Determina-se a localização dos zeros $\alpha_n < \alpha_{n-1} < \cdots < \alpha_2 < \alpha_1$.
2. Partindo de um valor $x^{(0)} > \alpha_1$, usando o método de Newton, calcula-se uma aproximação numérica para o maior zero α_1, com a precisão desejada.
3. Pelo algoritmo de Hörner/Ruffini divide-se $P_n(x)$ por $x - \alpha_1$ e regressa-se ao passo 2 para determinar α_2. Este processo é conhecido por deflacção. Repetindo sucessivamente este processo, determinamos numericamente todos os zeros do polinómio.
4. Para refinar as aproximações obtidas, aplica-se o método de Newton a P_n sendo as aproximaçẽs iniciais os valores obtidos no passo 3.

O método de Newton-Hörner pode ser descrito da seguinte forma. Dada uma estimativa inicial $r_j^{(0)}$ para a raiz α_j, calcular, para cada $k \geq 0$ até à convergência

$$r_j^{(k+1)} = r_j^{(k)} - \frac{P_n(r_j^{(k)})}{Q_{n-1}(r_j^{(k)}; r_j^{(k)})}.$$

Exercício 4.11 Construa o algoritmo implícito no procedimento descrito anteriormente.

No caso de alguma das raízes α ter multiplicidade $m > 1$ podemos escrever

$$P_n(x) = (x - \alpha)^m P_{n-m}(x),$$

onde o polinómio p_{n-m}, de grau $n - m$, é tal que $P_{n-m}(\alpha) \neq 0$. A aproximação desta raiz é feita com recurso ao método de Newton modificado

$$r^{(k+1)} = x^{(k)} - m\frac{P_n(r^{(k)})}{P_n'(r^{(k)})}, \qquad k = 0, 1, 2, \ldots. \tag{4.17}$$

Exercício 4.12 Prove que se α for um zero de multiplicidade m de um polinómio P, o método de Newton modificado (4.17) converge localmente (quais as condições de convergência?) e de forma quadrática para α.

Para calcular as raízes complexas de uma equação algébrica o método da bissecção não pode ser usado. Quanto ao método de Newton, ele só convergirá para uma raiz complexa se a aproximação inicial for um número complexo (e se forem satisfeitas as condições de convergência), sendo todo o processo realizado com aritmética complexa. Note-se que, após determinada uma raiz complexa, ficamos imediatamente a conhecer outra raiz (a sua conjugada).

4.7 Sistemas de equações não lineares

Nesta secção vamos descrever, de forma sucinta, a aplicação do método de Newton à resolução numérica de sistemas de equações não lineares.

Consideremos o ponto $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ e a aplicação F, suficientemente regular, definida por

$$\begin{array}{rccl} F\ : & \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ & (x_1, \ldots, x_n) & \longrightarrow & (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n)) \end{array}.$$

É nosso objectivo determinar a solução $\alpha = (\alpha_1, \ldots, \alpha_n)$ do sistema de n equações em n incógnitas

$$\left\{ \begin{array}{rcl} f_1(x_1, \ldots, x_n) & = & 0 \\ \vdots & & \vdots \\ f_n(x_1, \ldots, x_n) & = & 0 \end{array} \right. ,$$

que, noutra notação, pode ser escrito na forma

$$F(x) = 0, \tag{4.18}$$

com

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}, \qquad 0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.19}$$

A resolução de sistemas de equações não lineares por processos analíticos pode ser bastante difícil ou mesmo impossível. Nesse caso temos necessidade de recorrer a métodos numéricos no sentido de obter uma solução aproximada. Iremos considerar métodos iterativos da forma

$$x^{(k+1)} = \Phi(x^{(k)}), \qquad k = 0, 1, \ldots, \tag{4.20}$$

com

$$\Phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_n(x) \end{bmatrix}, \tag{4.21}$$

que determinem uma sucessão de aproximações para uma raiz α da equação vectorial (4.18), a partir de uma dada aproximação inicial

$$x^{(0)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^{(0)} . \tag{4.22}$$

Uma questão essencial quando se lida com métodos iterativos tem a ver com a convergência do processo: dada uma sucessão de aproximações $\{x^{(k)}\}$ gerada pelo processo iterativo, como saber se ela é convergente para a solução α do sistema?

Definição 4.5 (Convergência) *A* sucessão de vectores $\{x^{(k)}\}$ *diz-se* convergente *para* $\alpha \in \mathbb{R}^n$ *se, para todo o* $\epsilon > 0$*, existe uma ordem* k_0 *tal que, para todo o* $k > k_0$*, se tem* $\|x^{(k)} - \alpha\| < \epsilon$*. Nesse caso escreve-se* $\lim_{k \to +\infty} x^{(k)} = \alpha$.

De notar que a noção de convergência depende do conceito de norma. Uma vez que é possível considerar várias normas, uma questão legítima seria a de saber se é possível que uma sucessão de vectores convirja quando se considera uma determinada norma e divirja quando se considera outra. Para as normas mais usuais (dadas no Exercício 1.1) é possível demonstrar que se uma sucessão de vectores convergir segundo uma das normas ela também converge quando se considera outra qualquer. Por este facto diz-se que estamos em presença de normas equivalentes.

Consideremos agora o problema da definição de critérios de paragem para processos iterativos aplicados ao cálculo das raízes de sistemas de equações não lineares $F(x) = 0$. Seja $\{x^{(k)}\}$ a sucessão de aproximações gerada pelo processo iterativo convergente para a solução α do sistema. Os critérios de paragem mais frequentes são:

1. Critério do erro absoluto: $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$;
2. Critério do erro relativo: $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon \|x^{(k)}\|$;
3. Critério do valor da função: $\|F(x^{(k)})\| \leq \varepsilon_1$, com $\varepsilon_1 \ll \varepsilon$;
4. Critério do número máximo de iterações: $k = k_{max}$.

Antes de passarmos à definição dos processos iterativos vamos considerar o problema da determinação da aproximação inicial que, para sistemas de equações, pode ser um problema de difícil resolução. Na prática existem processos que permitem, *a priori*, determinar

boas estimativas iniciais para a solução pretendida. Esses processos dependem muito do problema em questão e como tal não são passíveis de um tratamento generalizado.

Existe, no entanto, uma forma de poder obter uma boa aproximação inicial quando os sistemas são de pequena dimensão. Essa forma é a localização gráfica. Este processo consiste na mera generalização do efectuado na secção anterior e, como tal, não iremos fazer a sua abordagem na forma geral mas sim recorrendo a um exemplo.

Exemplo 4.8 Considere-se o sistema de equações não lineares

$$\begin{cases} x^2+y^2 &= 1 \\ xy+x &= 1 \end{cases}.$$

Traçando o gráfico de $f_1(x,y)=0$ e $f_2(x,y)=0$, com

$$\begin{cases} f_1(x,y) &= x^2+y^2-1 \\ f_2(x,y) &= xy+x-1 \end{cases},$$

verificamos que uma solução do sistema é $(x,y)=(1,0)$ e que a outra está próxima de $(x,y)^{(0)}=(1,1)$.

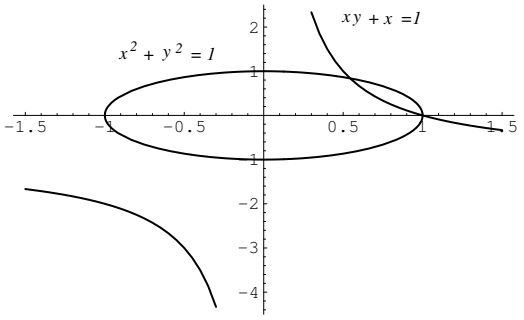

Figura 4.7: Localização gráfica.

Seja dado o sistema de equações não lineares $F(x)=0$ definido por (4.19). Pretendemos determinar uma aproximação para a raiz $\alpha=(\alpha_1,\ldots,\alpha_n)$ do referido sistema sendo dada uma aproximação inicial (4.22). Suponhamos que $F\in C^2(\mathcal{V}_\alpha)$, com $\mathcal{V}_\alpha$ uma vizinhança de α. Pela fórmula de Taylor temos que, se $x^{(0)}\in\mathcal{V}_\alpha$,

$$F(\alpha)=F(x^{(0)})+J_F(x^{(0)})(\alpha-x^{(0)})+\cdots,$$

onde

$$J_F(x)=\begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(x) & \cdots & \dfrac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(x) & \cdots & \dfrac{\partial f_n}{\partial x_n}(x) \end{bmatrix}$$

é a matriz de Jacobi de F no ponto x. Como $F(\alpha)=0$ e supondo $\det(J_F(x))\neq 0$, para todo o $x\in\mathcal{V}_\alpha$, podemos definir o processo iterativo

$$x^{(k+1)}=x^{(k)}-J_F^{-1}(x^{(k)})F(x^{(k)}),\qquad k=0,1\ldots,$$

que pretendemos que seja convergente para α. Este processo iterativo é designado por método de Newton ou método de Newton-Raphson.

Antes de apresentar o algoritmo que traduz o método de Newton, notemos que podemos evitar, em cada iteração, o cálculo da matriz inversa $J_F^{-1}(x^{(k)})$ se fizermos

$$\left\{ \begin{array}{rcl} J_F(x^{(k)})\delta x^{(k)} & = & -F(x^{(k)}) \\ \delta x^{(k)} & = & x^{(k+1)} - x^{(k)} \end{array} \right. . \tag{4.23}$$

Algoritmo 4.5 Método de Newton

Dados: $x^{(0)}$, ε e $k = k_{max}$
$x := x^{(0)}$
$k := 0$
$erro := \varepsilon + 1$
Enquanto $erro > \varepsilon$ e $k < k_{max}$ fazer
    Se $\det(J_F(x)) = 0$ então parar
    $k := k + 1$
    Resolver $J_F(x)\delta = -F(x)$
    $x := x + \delta$
    $erro := \|\delta\|$
Resultado: $\alpha \approx x$

Notemos que o carácter local da convergência deste método nos obriga a ter o cuidado de escolher uma aproximação inicial que esteja *suficientemente próxima* da solução que pertendemos determinar.

Exercício 4.13 Determine uma aproximação para a solução de

$$F(x) = 0 \Leftrightarrow \left\{ \begin{array}{rcl} x^2 + y^2 - 1 & = & 0 \\ xy + x - 1 & = & 0 \end{array} \right. ,$$

diferente de $(1, 0)$, efectuando duas iterações de método de Newton. Indique uma estimativa para o erro cometido.

Resolução: Seja $\alpha = (\alpha_1, \alpha_2)$ a solução pretendida. Como vimos, a aproximação inicial pode ser dada por $(x, y)^{(0)} = (1, 1)$.

Para não sobrecarregar a notação consideremos $(x, y)^{(k)} = (x_k, y_k)$, $k = 0, 1, \ldots$. Como

$$J_F(x_k, y_k) = \left[\begin{array}{cc} 2x_k & 2y_k \\ y_k + 1 & x_k \end{array} \right],$$

temos que

$$\det(J_F(x_k, y_k)) \neq 0 \Leftrightarrow x_k^2 - y_k^2 - y_k \neq 0.$$

Apliquemos o método de Newton na forma (4.23).

- Primeira iteração.
  Como $x_0^2 - y_0^2 - y_0 = -1 \neq 0$ podemos efectuar a primeira iteração do método. Assim

$$\begin{bmatrix} 2x_0 & 2y_0 \\ y_0+1 & x_0 \end{bmatrix} \begin{bmatrix} \delta x_0 \\ \delta y_0 \end{bmatrix} = - \begin{bmatrix} x_0^2+y_0^2-1 \\ x_0y_0+x_0-1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \delta x_0 \\ \delta y_0 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

  Daqui sai que

$$\begin{bmatrix} \delta x_0 \\ \delta y_0 \end{bmatrix} = \begin{bmatrix} -0{,}5 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0{,}5 \\ 1 \end{bmatrix}.$$

- Segunda iteração.
  Como $x_1^2 - y_1^2 - y_1 = -1{,}75 \neq 0$ podemos efectuar a segunda iteração do método. Assim obtemos, de modo análogo,

$$\begin{bmatrix} 1 & 2 \\ 2 & 0{,}5 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta y_1 \end{bmatrix} = \begin{bmatrix} -0{,}25 \\ 0 \end{bmatrix}.$$

  Daqui sai que

$$\begin{bmatrix} \delta x_1 \\ \delta y_1 \end{bmatrix} = \begin{bmatrix} 1/28 \\ -1/7 \end{bmatrix} \Rightarrow \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 15/28 \\ 6/7 \end{bmatrix}.$$

Temos assim que

$$\alpha \approx \left(\frac{15}{28}, \frac{6}{7}\right) = (0{,}5357, 0{,}8571),$$

sendo uma estimativa para o erro cometido dada por

$$\|\alpha - x^{(2)}\|_\infty \approx \|x^{(2)} - x^{(1)}\|_\infty = \max\left\{\frac{1}{28}, \frac{1}{7}\right\} = \frac{1}{7} = 0{,}1429.$$

4.8 Problemas

4.8.1 Exercícios para resolver nas aulas

Exercício 4.14 Seja $f : \mathbb{R} \longrightarrow \mathbb{R}$ definida por $f(x) = |x| - e^x$, $x \in \mathbb{R}$.

1. Mostre que a função f tem um único zero no intervalo $[-1, 0]$.
2. Usando o método da bissecção, determine um valor aproximado para o zero isolado na alínea anterior, com um erro absoluto que não exceda 0.15.

Exercício 4.15 Um avião em voo descreve uma trajetória em que a altitude, para $t \in [0, 1]$ (minutos), pode ser traduzida pela expressão

$$h(t) = (t-1)e^t - t + 3.$$

1. Seja $\alpha \in]0, 1[$ o instante para o qual o avião está mais perto do solo. Mostre que o método da bissecção, quando aplicado a uma função conveniente, no intervalo $[0, 1]$, gera uma sucessão que converge para α.
2. Efetue 4 iterações do método da bissecção para determinar uma aproximação para α e indique um majorante para o erro absoluto cometido.

Exercício 4.16 **(Matlab)** Use o método da bissecção para aproximar a solução, com erro inferior a 10^{-2}, da equação $x + 0{,}5 + 2\cos(\pi x) = 0$ no intervalo $[0{,}5, 1]$.

Exercício 4.17 **(Matlab)** Determine o número mínimo de iterações necessárias para aproximar, pelo método da bissecção e com uma precisão de 10^{-1}, a solução de $x^3 - x - 1 = 0$ no intervalo $[1, 2]$. Determine tal aproximação com a precisão indicada.

Exercício 4.18 **(Matlab)** A componente forçada de uma tensão transitória para um dado circuito pode ser traduzida pela expressão

$$\mathcal{E}(t) = e^{-0{,}06\pi t}\sin(2t - \pi).$$

Usando o método da bissecção duas vezes, determine a tensão máxima deste circuito.

Exercício 4.19 **(Matlab)** Dada a função

$$f(x) = \cosh x + \cos x - \gamma,$$

para $\gamma = 1, 2, 3$, averigúe, em cada caso, se f tem zeros e, caso existam, aproxime o seu valor pelo método da bissecção, com um erro inferior a 10^{-10}.

Exercício 4.20 **(Matlab)** A equação de um gás é definida por

$$\left[p + a(N/V)^2\right](V - Nb) = kNT,$$

em que p é a pressão, V o volume ocupado pelo gás à temperatura T, a e b são dois coeficientes que dependem do gás considerado, N é o número de moléculas contidas no volume V e k é a constante de Boltzmann ($k = 1{,}3806503 \times 10^{-23}$). Aproxime o volume V ocupado por 1000 moléculas de dióxido de carbono à temperatura $T = 300$ K e pressão $p = 3{,}5 \times 10^7$ Pa, pelo método da bissecção, com um erro inferior a 10^{-12}. Para o dióxido de carbono, tem-se $a = 0{,}401$ Pa, $b = 42{,}7 \times 10^{-6}$ m^3.

Exercício 4.21 Considere a equação $f(x) = e^{2x} - e^{-x} - 2$.

1. Determine um intervalo de amplitude 1 que contenha o zero da função f.
2. Mostre que pode aproximar o zero localizado na alínea pelo método de Newton.
3. Obtenha uma aproximação fazendo duas iterações do método de Newton e indique uma estimativa para o erro absoluto cometido na aproximação.

Exercício 4.22 Considere a equação $e^{1/x} - x = 0$.

1. Mostre que a equação anterior tem uma única raiz no intervalo [1,2].
2. Considere a aproximação inicial $x^{(0)} = 1$. Verifique o método de Newton converge.
3. Aproxime a raiz fazendo duas iterações do método de Newton usando a aproximação inicial anterior.

Exercício 4.23 Considere a equação

$$e^x - 3x^2 = 0,$$

que tem três raízes reais $\alpha_1 < \alpha_2 < \alpha_3$ tais que $\alpha_1 \in [-0{,}6,\ -0{,}4]$, $\alpha_2 \in [0{,}8,\ 1{,}0]$, $\alpha_3 \in [3{,}6,\ 3{,}8]$. Mostre que o método de Newton, escolhendo convenientemente a aproximação inicial $x^{(0)} \in [3{,}6,\ 3{,}8]$, converge para a raiz α_3 e utilize este método para obter um valor aproximado $x^{(k)}$ de α_3 tal que $|x^{(k)} - x^{(k-1)}| < 10^{-3}$.

Exercício 4.24 Considere a função f definida por $f(x) = e^{-x}\ln x$, $x > 0$. Utilizando o método de Newton, aproxime a abcissa do seu ponto de inflexão, em segunda aproximação, partindo de um intervalo com amplitude inferior ou igual a 1. Indique uma estimativa para o erro cometido.

Exercício 4.25 Use o método de Newton para aproximar, com erro inferior a 10^{-4}, o valor de x correspondente ao ponto do gráfico de $y = x^2$ mais próximo de $(1, 0)$.

Exercício 4.26 Aplicar o método de Newton ao cálculo da raiz quadrada de um número positivo a. Proceder de maneira análoga para calcular a raiz cúbica de a.

Exercício 4.27 (Matlab) Um projéctil é lançado com uma velocidade v_0 e um ângulo α num túnel de altura h e atinge o seu máximo quando α for tal que

$$\sin(\alpha) = \sqrt{2gh/v_0^2},$$

onde $g = 9{,}8$ m/s^2 é a aceleração da gravidade. Calcular α utilizando o método de Newton, supondo que $v_0 = 10$ m/s e $h = 1$ m.

Exercício 4.28 (Matlab) Considere o sistema mecânico representado por quatro barras rígidas a_i, i=1,2,3,4. Para qualquer valor admissível do ângulo β formado pelas barras a_1 e a_4, determinamos o valor do ângulo correspondente α formado pelas barras a_1 e a_2. Partindo da identidade vectorial $a_1 - a_2 - a_3 - a_4 = 0$ e observando que a barra a_1 está sempre alinhada com o eixo dos x, podemos obter a seguinte relação entre α e β

$$\frac{a_1}{a_2}\cos\beta - \frac{a_1}{a_4}\cos\alpha - \cos(\beta - \alpha) = -\frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4},$$

onde a_i é o comprimento da i-ésima barra. Esta igualdade chama-se *equação de Freudenstein*, e pode escrever-se do seguinte modo: $f(\alpha) = 0$, em que

$$f(x) = \frac{a_1}{a_2}\cos\beta - \frac{a_1}{a_4}\cos x - \cos(\beta - x) + \frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4}.$$

Só para valores especiais de β é que existe uma expressão explícita da solução. Refira-se ainda que não existe solução para todos os valores de β, e que a solução existindo poderá não ser única. A fim de resolver a equação para qualquer valor de β a variar entre 0 e π, deveremos recorrer a métodos numéricos. Aproxime o valor de α recorrendo ao método de Newton, com $\beta \in [0, \frac{2\pi}{3}]$ e com uma tolerância de 10^{-5}. Suponha que os comprimentos das barras são, respectivamente, $a_1 = 10$ cm, $a_2 = 13$ cm, $a_3 = 8$cm e $a_4 = 10$ cm. Para cada valor de β considerar dois dados iniciais $x^{(0)} = -0{,}1$ e $x^{(0)} = \frac{2\pi}{3}$.

Exercício 4.29 (Matlab) Considere a função $f(x) = e^x - 2x^2$.

1. Localize as raízes da equação $f(x) = 0$.
2. Determine uma estimativa para a maior raiz, usando o método de Newton, com 5, 10 e 15 iterações.
3. Repita a alínea anterior, recorrendo ao método da bissecção. Compare os resultados obtidos.

Exercício 4.30 Mostre que

$$x = \frac{1}{2}\cos x$$

tem uma única solução α. Obtenha, em seguida, um intervalo $[a, b]$ que contenha α e tal que, para todo o $x^{(0)}$ nesse intervalo, a sucessão

$$x^{(n+1)} = \frac{1}{2}\cos x^{(n)}, \qquad n = 0, 1, 2, \ldots,$$

convirja para α. Justifique.

Exercício 4.31 Considere a equação $3x^2 - e^x = 0$.

1. Mostre que a equação anterior tem apenas uma raiz real negativa e determine um intervalo com amplitude um, I, que a contenha.

2. Considere o método iterativo definido por $x^{(k+1)} = \phi(x^{(k)})$, com

   $$\phi(x) = -\sqrt{\frac{e^x}{3}}.$$

   Mostre que, qualquer que seja a aproximação inicial $x^{(0)} \in I$, o método converge para a raiz anterior.

3. Determine o número de iterações que deverá efetuar com o método dado na alínea anterior por forma a obter uma aproximação para a raiz da equação com uma estimativa para o erro absoluto que não exceda $0{,}5 \times 10^{-3}$.

Exercício 4.32 O número de ouro ϕ, pode ser obtido como solução positiva da equação

$$x = 1 + \frac{1}{x}.$$

1. Mostre que a equação anterior tem uma única raiz positiva em $[1{,}5,\ 2]$ e conclua que $\phi \in [1{,}5,\ 2]$.

2. Pretende-se determinar um valor aproximado para ϕ usando o método do ponto fixo $x_{n+1} = 1 + \frac{1}{x_n}$. Mostre que o método anterior converge para ϕ, qualquer que seja $x^{(0)} \in [1{,}5,\ 2]$.

3. Determine quantas iterações são necessárias efetuar para obter uma aproximação para ϕ com duas casas decimais corretas.

Exercício 4.33 Determine os extremos locais da função $f(x) = \frac{x^3}{3} + 10\sin x$, com um erro inferior a 10^{-4}, usando o método iterativo do ponto fixo.

Exercício 4.34 Determine uma aproximação para a maior raiz de $e^x - 4x^2 = 0$, usando o método do ponto fixo. Indique um majorante do erro da aproximação obtida.

Exercício 4.35 Seja ϕ_N a função iteradora do método de Newton considerado como uma iteração do ponto fixo. Mostre que

$$\phi_N'(\alpha) = 1 - \frac{1}{m},$$

onde α é um zero de f de multiplicidade m. Deduzir que o método de Newton converge quadraticamente se α for uma raiz simples de $f(x) = 0$ e linearmente no caso contrário.

Exercício 4.36 **(Matlab)** Considere a equação $e^x - x - 2 = 0$.

1. Verifique que a equação dada tem uma única solução no intervalo $[1, 2]$.
2. Para aproximar o valor da solução pretende-se utilizar o método do ponto fixo, com uma das seguintes funções iteradoras:

   $$\phi_1(x) = e^x - 2, \qquad \phi_2(x) = \ln(x + 2) \qquad \text{e} \qquad \phi_3(x) = x - 0{,}1(e^x - x - 2).$$

   Indique qual ou quais das funções garante a convergência do método para a referida solução. Escolha uma aproximação inicial aproxime a solução da equação com um número de iterações suficientes por forma a que o valor absoluto da diferença entre duas iterações consecutivas não exceda 10^{-4}.
3. Recorra ao método de Newton para aproximar a mesma solução e compare com o resultado obtido na alínea anterior.

Exercício 4.37 **(Matlab)** Considere a equação $\ln x - \frac{1}{x} = 0$.

1. Verifique que a equação dada tem uma única solução no intervalo $[1, 2]$.
2. Para aproximar o valor da solução pretende-se utilizar o método do ponto fixo, com uma das seguintes funções iteradoras:

   $$\phi_1(x) = e^{\frac{1}{x}} \qquad \text{e} \qquad \phi_2(x) = 1/\ln(x).$$

   Indique qual das funções garante a convergência do método para a referida solução, qualquer que seja o ponto inicial. Escolha uma aproximação inicial e aproxime a solução da equação com um número de iterações suficientes por forma a que o valor absoluto da diferença entre duas iterações consecutivas não exceda 10^{-4}
3. Recorra ao método da bissecção para aproximar a mesma solução e compare com o resultado obtido na alínea anterior.

Exercício 4.38 Considere a equação $x^6 - 4x^5 + 6x^4 - 2x^3 - kx^2 + mx - n = 0$.

1. Determine m, n e k de modo a que esta equação admita 1 como raiz tripla.
2. Localize as raízes da equação e determine a menor delas pelo método que considerar mais conveniente.

Exercício 4.39 Considere a equação polinomial $x^3 + kx^2 + 2x - 1 = 0$, com $k \geq 0$.

1. Determine o conjunto de todos os valores de k para os quais a equação tem uma única raiz no intervalo $[0, 1]$.
2. Tome para k o menor valor inteiro positivo do conjunto obtido na alínea anterior e faça a separação completa das raízes da equação dada.
3. Aproxime a menor raiz real daquela equação pelo método de Newton.

Exercício 4.40 **(Matlab)** Considere o polinómio $P(x) = x^3 - 36x^2 + 188x - 240$.

1. Verifique que um dos zeros de $P(x)$ se localiza no intervalo $[29{,}5, 31]$.
2. Para aproximar o zero referido na alínea anterior, podem ser usadas diferentes estratégias como, por exemplo:
   (a) método da bissecção no referido intervalo, com tol$= 10^{-4}$;
   (b) método de Newton, com a aproximação inicial $x^{(0)} = 29{,}5$ e tol$= 10^{-4}$.

   Compare os resultados obtidos com os dois métodos.
3. Para aproximar o mesmo zero de $P(x)$ use o método do ponto fixo com a função iteradora

$$g(x) = (240 + 36x^2 - x^3)/188.$$

   Considere as seguintes aproximações iniciais $x^{(0)} = 1$, $x^{(0)} = 3$ e $x^{(0)} = 31$ e compare os resultados obtidos, fazendo tol$= 10^{-4}$.

Exercício 4.41 **(Matlab)** Determine todas as raízes reais de $x^3 + x^2 + 2x - 1 = 0$ e $-x^3 - x^2 + 3x + 1 = 0$ usando o algoritmo de Newton-Hörner.

Exercício 4.42 Considere a seguinte função $F\colon \mathbb{R}^2 \to \mathbb{R}^2$:

$$F(x) = \begin{bmatrix} x_1^2 + x_2^2 - 2 \\ \exp(x_1 - 1) + x_2^3 - 2 \end{bmatrix}.$$

1. Identifique uma das raízes, α, de $F(x) = 0$.
2. Efectue uma iteração do Método de Newton para aproximar α a partir do ponto inicial $x^{(0)} = (2, 1/2)$.

Exercício 4.43 Determine uma aproximação para a solução de

$$\begin{cases} x^2 + y^2 &= 4 \\ x^2 - y^2 &= 1 \end{cases}$$

localizada no segundo quadrante, efetuando duas iterações do método de Newton.

Exercício 4.44 Aplique duas iterações do método de Newton ao sistema de equações não lineares

$$\begin{cases} e^{x_1} - 1 &= 0 \\ e^{x_2} - 1 &= 0 \end{cases},$$

começando com $x^{(0)} = (-10, -10)$. O que é que aconteceria se continuasse a aplicar o método de Newton?

Exercício 4.45 **(Matlab)** O seguinte sistema de equações não lineares possui duas soluções, uma das quais é $(x, y) = (3, 0)$,

$$\begin{cases} x^2 - 2x + y^2 - 3 &= 0 \\ x(6 - x) + y - 9 &= 0 \end{cases}.$$

Localize e determine a outra solução efectuando cinco iterações do método de Newton.

Exercício 4.46 **(Matlab)** Considere o sistema de equações não lineares

$$\begin{cases} x_1^2 + x_2^2 &= 1 \\ \sin(\frac{\pi x_1}{2}) + x_2^3 &= 0 \end{cases}.$$

1. Localize graficamente as duas soluções do sistema.
2. Efectue duas iterações do método de Newton, partindo da aproximação inicial $x^{(0)} = (1, 1)$.
3. Repita a alínea anterior com $x^{(0)} = (-1, -1)$. O que pode concluir?

Exercício 4.47 **(Matlab)** Recorra ao método de Newton para determinar uma solução do sistema

$$\begin{cases} x = 0{,}7 \sin x + 0{,}2 \cos y \\ y = 0{,}7 \cos x - 0{,}2 \sin y \end{cases}$$

próxima de $(0{,}5, 0{,}5)$.

Exercício 4.48 **(Matlab)** Pretende construir-se uma ponte entre duas margens de um rio que, por razões económicas, seja o mais curta possível. Sabendo que, na região onde se pretende construir a ponte, as margens do rio têm a forma das curvas $y = e^x$ e $y = \ln x$, determine um valor aproximado do comprimento da ponte.

Exercício 4.49 Efectue duas iterações do método de Newton para calcular o mínimo local da função quadrática

$$f(x_1, x_2) = x_1^2 - 4x_1x_2 + x_2^2.$$

4.8.2 Exercícios de aplicação à engenharia

Exercício 4.50 Para determinar a queda de pressão em escoamentos de líquidos em tubos cilíndricos, torna-se necessário obter o chamado factor de atrito, f, que é dado pela relação empírica

$$\frac{1}{\sqrt{f}} = \frac{1}{k} \ln\left(Re\sqrt{f}\right) + \left(14 - \frac{5{,}6}{k}\right),$$

em que k é a rugosidade e Re o número (adimensional) de Reynolds do escoamento. Para um valor de $k = 0{,}28$ e um número de Reynolds $Re = 3750$, determine o valor de f.

Exercício 4.51 Baseado no trabalho de Frank-Kamenetski (1955), as temperaturas no interior de um material com fontes de calor embebidas podem ser determinadas pela equação

$$e^{-0{,}5t} \cosh\left(e^{0{,}5t}\right) = \sqrt{0{,}5L_{er}}.$$

Dado $L_{er} = 0{,}088$, determine t.

Exercício 4.52 A concentração, C, de uma bactéria poluente num lago decresce de acordo com a expressão

$$C = 80e^{-2t} + 20e^{-0,1t},$$

onde t representa o tempo. Determine o tempo necessário para que a concentração de bactérias fique reduzida a 10.

Exercício 4.53 Um medicamento administrado a um doente produz uma concentração na corrente sanguínea dada por

$$c(t) = Ate^{-t/3} \text{ mg/ml},$$

t horas depois de injectadas A unidades. A concentração máxima de segurança é de 1 mg/ml.

1. Que quantidade deve ser injectada para que seja atingida a concentração máxima de segurança e em que altura ocorre esse máximo?

2. Uma concentração adicional do mesmo medicamento deve ser administrada no doente depois da concentração ter descido para $0{,}25$ mg/ml. Determine quando é que a segunda injecção deve ser administrada (em minutos).

3. Assumindo que a concentração após injecções consecutivas é aditiva e que $0{,}75\%$ da quantidade original injectada é administrada na segunda injecção, em que altura deve ser dada a terceira injecção?

Exercício 4.54 Em engenharia ambiental, a equação que se segue pode ser usada para calcular o nível de oxigénio, c, existente num rio a jusante de um local de descarga de esgoto,

$$c = 10 - 15(e^{-0,1x} - e^{-0,5x}),$$

em que x representa a distância a partir do local de descarga. Usando um método à sua escolha, determine o local (a partir da descarga) em que o nível de oxigénio atinge o valor 4.

Sugestão: Sabe-se que o referido local se encontra, no máximo, a 5 km a jusante do local de descarga.

Exercício 4.55 Em engenharia oceânica, a equação para a altura de uma determinada onda num cais é dada por

$$h = h_0 \left(\sin\left(\frac{2\pi x}{\lambda}\right) \cos\left(\frac{2\pi t v}{\lambda}\right) + e^{-x} \right).$$

Determine uma aproximação para o valor de x sabendo que $h = 0{,}5h_0$, $\lambda = 20$, $t = 10$ e $v = 50$.

Exercício 4.56 Num escoamento com superfície livre pode definir-se uma camada junto ao fundo (designada por camada limite) onde as características do escoamento são significativamente diferentes das que se verificam acima dessa camada. Pode provar-se que a espessura da camada limite é $\delta = 5z$, sendo z, para um escoamento com determinadas características, dado por $|z| \log |75z| = 2$.

1. Localize graficamente as raízes reais desta equação.

2. Determine a segunda aproximação dada pelo método de Newton para a espessura da camada limite, δ.

3. Indique um limite superior para o erro cometido na aproximação obtida na alínea anterior.

Exercício 4.57 De Santis (1976) deduziu uma relação para o factor de compressibilidade dos gases reais da forma

$$z = \frac{1 + y + y^2 - y^3}{(1-y)^3},$$

onde $y = b/4\nu$, sendo b a correcção de van der Waals e ν o volume molar. Se $z = 0{,}892$ qual o valor de y?

Exercício 4.58 Pretende construir-se um depósito semi-esférico, de raio r, para armazenar um líquido até uma altura h. Sabendo que o volume do referido líquido é dado pela expressão

$$V = \frac{\pi\left(2r^3 - 3r^2h + h^3\right)}{3},$$

qual o raio com que se deve construir o depósito se se pretender guardar no máximo 250 litros de líquido a uma altura de 2 metros?

Exercício 4.59 Um corpo de massa 1 kg, que se move apenas ao longo de uma linha recta e que inicialmente se encontra em repouso no ponto de coordenadas $x = 2$, fica sujeito a uma força cuja intensidade em cada instante t é dada por

$$F(t) = -1 + 2t - 3t^2.$$

Localize e separe os instantes de tempo em que o corpo passa pela origem do referencial.

Capítulo 5

Interpolação

Seja f uma função real definida num conjunto de pontos $x_0, x_1, \ldots, x_n$. Pretende-se calcular o valor de $f(\overline{x})$, com $\overline{x} \neq x_i$, $i = 0, 1, \ldots, n$. Tal situação é muito frequente, por exemplo, no contexto das equações diferenciais. Quando se usam métodos numéricos para aproximar a solução de uma equação diferencial esta fica apenas conhecida num conjunto de pontos. A interpolação permite assim encontrar uma função que passa por esse conjunto de pontos e que pode funcionar como uma aproximação à solução da equação.

Em linhas gerais, o conceito de interpolação consiste em determinar uma função $\psi(x) = a_0\psi_0(x) + \cdots + a_n\psi_n(x)$, gerada por uma certa família de funções $\{\psi_k\}_{k=0}^n$, por forma a que

$$f(x_i) = \psi(x_i), \qquad i = 0, 1, \ldots, n.$$

A função ψ nestas condições é designada por função interpoladora de f nos pontos de suporte (interpolação) $x_0, x_1, \ldots, x_n$.

Nada nos garante que o problema da interpolação tenha sempre solução. Por exemplo, fazendo $\psi_0(x) = 1$ e $\psi_1(x) = x^2$, não existe nenhuma função $\psi(x) = a_0 + a_1x^2$ que passe nos pontos $(1, 1)$ e $(-1, 0)$.

5.1 Interpolação polinomial de Lagrange

Um caso particular de interpolação com grande importância devido ao grande número de aplicações é a interpolação polinomial. Além disso, as fórmulas desenvolvidas para a interpolação polinomial estão na base do desenvolvimento de muitos métodos numéricos para o cálculo de raízes de equações não lineares, cálculo de integrais e derivadas, bem como a resolução de equações diferenciais.

No caso da interpolação polinomial, as funções geradoras são, por exemplo, $\psi_k(x) = x^k$, $k = 0, 1, \ldots, n$.

Definição 5.1 *Seja f uma função definida num intervalo $[a, b]$ e conhecida nos pontos da partição*

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b. \tag{5.1}$$

Um polinómio P que satisfaz

$$f(x_i) = P(x_i), \qquad i = 0, 1, \ldots, n, \tag{5.2}$$

é chamado polinómio interpolador (de Lagrange) *de f nos pontos da partição dada.*

Exercício 5.1 Dada a tabela

x_i	2,3	2,4	2,5	2,6
$\log x_i$	0,361728	0,380211	0,397940	0,414973

,

determine o valor aproximado de $\log 2{,}45$, usando interpolação polinomial.

Resolução: Vamos calcular o polinómio P_3 de grau menor ou igual a 3, interpolador de $y = \log x$ nos pontos 2,3, 2,4, 2,5 e 2,6. De acordo com a definição temos $P_3(2{,}3) = 0{,}361728$, $P_3(2{,}4) = 0{,}380211$, $P_3(2{,}5) = 0{,}397940$, e $P_3(2{,}6) = 0{,}414973$. Isto é, se $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, temos que

$$\begin{cases} a_0 + 2{,}3a_1 + 5{,}29a_2 + 12{,}167a_3 = 0{,}361728 \\ a_0 + 2{,}4a_1 + 5{,}76a_2 + 13{,}824a_3 = 0{,}380211 \\ a_0 + 2{,}5a_1 + 6{,}25a_2 + 15{,}625a_3 = 0{,}397940 \\ a_0 + 2{,}6a_1 + 6{,}76a_2 + 17{,}576a_3 = 0{,}414973 \end{cases}.$$

Sendo o sistema possível e determinado tal polinómio existe e é único. Assim

$$P_3(x) = -0{,}404885 + 0{,}528963x - 0{,}107300x^2 + 0{,}009667x^3$$

é o polinómio pretendido. Temos então que $\log 2{,}45 \approx P_3(2{,}45) = 0{,}389170$. Compare-se este valor com o valor exacto $\log 2{,}45 = 0{,}38916608\ldots$. Note-se que o erro cometido na aproximação não excede $0{,}4 \times 10^{-5}$.

A determinação do polinómio interpolador por este processo é pouco eficiente e pouco estável. Quanto à eficiência, note-se que a resolução do sistema linear requer $(n+1)^3/3 + (n+1)^2 - (n+1)/3$ multiplicações/adições ($\mathcal{O}(n^3)$ operações). Para além de pouco eficiente, este processo também é pouco estável: na prática verifica-se que este método não permite ir além de valores de n da ordem da dezena quando se trabalha em aritmética com 6 ou 7 decimais.

5.1.1 Existência e unicidade. Fórmula de Lagrange

O método de determinar um polinómio interpolador usado no exercício anterior não é eficiente nem estável. Apresentaremos, neste capítulo, alguns métodos mais eficientes para a sua determinação.

O próximo teorema estabelece a existência e unicidade do polinómio de grau inferior ou igual a n interpolador de uma função em $n+1$ pontos distintos. Além disso, indica-nos um processo que permite a sua determinação.

Teorema 5.1 (Lagrange) *Seja f uma função definida num intervalo $[a, b]$ e conhecida nos pontos da partição (5.1). Existe um e um só polinómio P_n de grau menor ou igual a n interpolador de f nos pontos dados.*

Demonstração: Consideremos o polinómio P_n definido por

$$P_n(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x), \tag{5.3}$$

em que

$$\ell_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \qquad i = 1, \ldots, n. \tag{5.4}$$

Notemos que cada ℓ_i é um polinómio de grau n. Além disso

$$\ell_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases},$$

isto é $\ell_i(x_j) = \delta_{i,j}$ onde $\delta_{i,j}$ representa o símbolo de Kronecker, em honra de Leopold Kronecker (1823-1891). Portanto a função P_n é um polinómio de grau menor ou igual a n que verifica as condições de interpolação (5.2), o que prova a existência de solução do problema em causa.

Para provar a unicidade, suponhamos que P_n e Q_n são dois polinómios de grau menor ou igual a n interpoladores de f nos pontos da partição dada. Então o polinómio

$$R_n(x) = P_n(x) - Q_n(x)$$

anula-se, pelo menos, nos pontos x_i, $i = 0, 1, \ldots, n$. Como R_n é um polinómio de grau menor ou igual a n, ele só pode ter $n + 1$ zeros se for identicamente nulo. Assim sendo, $P_n(x) = Q_n(x)$, para todo o x, o que prova o pretendido. □

As expressões (5.3) e (5.4) definem a fórmula de Lagrange para calcular o polinómio interpolador de f nos pontos (5.1).

O resultado anterior diz-nos que por $n+1$ pontos passa um e um só polinómio de grau n. Assim temos que, se a funçao f a interpolar for um polinómio de grau n coincide com o seu polinómio interpolador do mesmo grau, podendo assim ser escrita na forma

$$f(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x).$$

Exercício 5.2 Dada a tabela

x_i	1	2	3	4
$f(x_i)$	4	15	40	85

,

determine uma aproximação para $f(1.5)$, usando interpolação cúbica.

Resolução: Temos que

$$\begin{aligned}
\ell_0(x) &= \frac{(x-2)(x-3)(x-4)}{(-1)(-2)(-3)} = -\frac{1}{6}(x-2)(x-3)(x-4), \\
\ell_1(x) &= \frac{(x-1)(x-3)(x-4)}{(1)(-1)(-2)} = \frac{1}{2}(x-1)(x-3)(x-4), \\
\ell_2(x) &= \frac{(x-1)(x-2)(x-4)}{(2)(1)(-1)} = -\frac{1}{2}(x-1)(x-2)(x-4), \\
\ell_3(x) &= \frac{(x-1)(x-2)(x-3)}{(3)(2)(1)} = \frac{1}{6}(x-1)(x-2)(x-3).
\end{aligned}$$

Assim

$$P_3(x) = \sum_{i=0}^{3} f(x_i)\ell_i(x) = \cdots = 1 + x + x^2 + x^3.$$

Atendendo à fórmula de Lagrange podemos construir o seguinte algoritmo para calcular o valor de $P_n(\overline{x})$, sendo P_n o polinómio interpolador de f nos $n+1$ pontos distintos $x_0, x_1, \ldots, x_n$.

Algoritmo 5.1 Fórmula de Lagrange

Dados: x_i, $i = 0, 1, \ldots, n$ e $\overline{x}$
$P := 0$
Para i de 0 até n fazer
    $\ell := 1$
    Para j de 0 até n fazer
        Se $j \neq i$ então $\ell := \ell(\overline{x} - x_j)/(x_i - x_j)$
    $P := P + f(x_i)\ell$
Resultado: $P_n(\overline{x}) = P$

Exercício 5.3 Mostre que fórmula de Lagrange pode ser escrita na forma

$$P_n(x) = \sum_{i=0}^{n} f(x_i) \frac{w(x)}{(x - x_i)w'(x_i)}, \tag{5.5}$$

sendo

$$w(x) = \prod_{j=0}^{n} (x - x_j). \tag{5.6}$$

Resolução: Atendendo a (5.6) temos que

$$w'(x) = \sum_{i=0}^{n} \prod_{j=0, j\neq i}^{n} (x - x_j) \Rightarrow w'(x_i) = \prod_{j=0, j\neq i}^{n} (x_i - x_j),$$

e como tal

$$\ell_i(x) = \frac{w(x)}{(x - x_i)w'(x_i)},$$

o que prova o pretendido.

Para determinar o esforço computacional necessário à obtenção do polinómio interpolador pela fórmula de Lagrange, note-se que, supondo as constantes

$$F_i = \frac{f(x_i)}{w'(x_i)}, \qquad i = 0, \ldots, n,$$

calculadas *a priori*, o cálculo do valor do polinómio interpolador num determinado ponto pode ser dado por

$$P_n(x) = w(x) \left[\frac{F_0}{x - x_0} + \cdots + \frac{F_n}{x - x_n} \right].$$

Este cálculo requer $n(n+1)$ multiplicações e $n(n+2)$ adições, isto é, $\mathcal{O}(n^2)$ operações, o que torna a fórmula de Lagrange muito mais eficiente que o processo matricial.

A fórmula de Lagrange possui, no entanto, o inconveniente de obrigar a refazer os cálculos dos polinómios (5.4) sempre que ocorra uma alteração nos pontos de suporte. Na prática esta situação acontece com frequência, por exemplo, quando pretendemos passar de p_n a p_{n+1}, pela adição de mais um ponto x_{n+1} ao suporte de interpolação, a fim de estudar o comportamento do erro. (Este problema é resolvido pelo algoritmo de Newton das diferenças divididas, que não será objecto de estudo nesta disciplina.)

5.1.2 Erro de interpolação

Por definição, o polinómio interpolador coincide com a função num dado conjunto de pontos de suporte. Interessa-nos saber, no entanto, se para os outros pontos do domínio da função, o polinómio interpolador constitui uma boa ou uma má aproximação para a função. Nesse sentido temos o seguinte teorema, que apresentamos sem demonstração.

Teorema 5.2 *Seja P_n o polinómio de grau menor ou igual a n interpolador da função f nos pontos da partição (5.1). Se $f \in C^n([a,b])$ e se $f^{(n+1)}$ for contínua em $]a,b[$, então para cada $x \in [a,b]$ existe $\xi = \xi(x) \in]a,b[$ tal que*

$$e(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x), \tag{5.7}$$

onde $w(x)$ é a função dada por (5.6).

Na prática, o erro de interpolação num ponto $\overline{x}$ é usado na forma

$$|e(\overline{x})| = |f(\overline{x}) - P_n(\overline{x})| \leq \frac{M_{n+1}}{(n+1)!} |w(\overline{x})|, \tag{5.8}$$

onde

$$M_{n+1} = \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right|.$$

Note-se a semelhança existente entre a fórmula do erro na interpolação e na fórmula de Taylor. A diferença está que, enquanto a primeira usa informação em vários pontos distintos, a segunda recorre apenas a um único ponto.

Exercício 5.4 Determine uma estimativa para o erro que se cometeu na aproximação efectuada no Exercício 5.1.

Resolução: Neste caso temos, atendendo a (5.8),

$$|e_3(\overline{x})| = |\log \overline{x} - P_3(\overline{x})| \leq \frac{M_4}{4!} |(\overline{x} - 2{,}3)(\overline{x} - 2{,}4)(\overline{x} - 2{,}5)(\overline{x} - 2{,}6)|,$$

onde

$$M_4 = \max_{x \in [2{,}3;2{,}6]} \left| f^{(4)}(x) \right| = \max_{x \in [2{,}3;2{,}6]} \frac{6}{x^4 \ln 10} = 0{,}093116.$$

Fazendo $\overline{x} = 2{,}45$ vem

$$|\log 2{,}45 - P_3(2.45)| \leq \frac{0{,}093116}{24} |(2{,}45 - 2{,}3)(2{,}45 - 2{,}4)(2{,}45 - 2{,}5)(2{,}45 - 2{,}6)|,$$

ou seja $|e_3(2{,}45)| \leq 0{,}917 \times 10^{-5}$.

Exercício 5.5 Considere a função $f(x) = \cos x$ para x em $[0, \pi]$. Determine o número de pontos a considerar no intervalo dado para que o erro máximo da aproximação de $f(x)$ por um polinómio interpolador nesses pontos seja inferior a $0{,}5$.

Resolução: Temos que, para $x \in [0, \pi]$,

$$|f(x) - P_n(x)| \leq \frac{\max\limits_{x\in[0,\pi]} \left|f^{(n+1)}(x)\right|}{(n+1)!}|w(x)| = \frac{|w(x)|}{(n+1)!} \leq \frac{\pi^{n+1}}{(n+1)!}.$$

Resta assim determinar qual o menor valor de n que satisfaz

$$\frac{\pi^{n+1}}{(n+1)!} \leq 0{,}5.$$

Por tentativas...

$$n = 6 \quad \Rightarrow \quad \frac{\pi^7}{7!} = 0{,}599$$

$$n = 7 \quad \Rightarrow \quad \frac{\pi^8}{8!} = 0{,}235.$$

Logo, 8 é o menor número de pontos que garante a aproximação pretendida.

Exercício 5.6 Seja f uma função nas condições do teorema anterior e tal que (5.8) se verifica. Seja P_n o seu polinómio interpolador nos pontos da partição (5.1).

1. Mostre que o seu erro de interpolação verifica

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{4(n+1)}h^{n+1}, \qquad \forall x \in [a, b], \tag{5.9}$$

com $h = \max\limits_{i=1,\ldots,n}(x_i - x_{i-1})$.

2. Mostre que se a partição (5.1) for uniforme se tem

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{4(n+1)n^{n+1}}(b-a)^{n+1}, \qquad \forall x \in [a, b].$$

Resolução: Vamos apenas demonstrar (5.9). Para tal, basta provar que

$$\max_{x\in[a,b]} |w(x)| \leq \frac{h^{n+1}n!}{4},$$

com w a função nodal (5.6). Vamos efectuar a demonstração por indução.

Para $n = 1$ temos que $w(x) = (x - x_0)(x - x_1)$. Assim, temos que

$$w'(x) = 0 \Rightarrow x = \frac{x_0 + x_1}{2}.$$

Como tal,

$$\max_{x\in[a,b]} |w(x)| = \max\left\{|w(a)|, \left|w\left(\frac{x_0+x_1}{2}\right)\right|, |w(b)|\right\} = \left|w\left(\frac{x_0+x_1}{2}\right)\right| \leq \frac{h^2}{4}.$$

Suponhamos que (5.9) se verifica para n e provemos a sua veracidade para $n+1$, isto é, que

$$\max_{x\in[a,b]}\left|\prod_{j=0}^{n+1}(x-x_j)\right| \leq \frac{h^{n+2}(n+1)!}{4},$$

com $a = x_0$ e $x_{n+1} = b$. Dado $x \in [a,b]$ temos que $x \in [a,x_n]$ ou $x \in [x_n,b]$. Consideremos a primeira hipótese. Temos então

$$\max_{x\in[a,b]}\left|\prod_{j=0}^{n+1}(x-x_j)\right| = \max_{x\in[a,b]}\left|\prod_{j=0}^{n}(x-x_j)\right||x-b| \leq \frac{h^{n+1}n!}{4}(n+1)h = \frac{h^{n+2}(n+1)!}{4},$$

o que prova o pretendido. O caso em que se considera a segunda hipótese demonstra-se de forma análoga.

Consideremos uma função f definida num intervalo $[a,b]$ onde está definida uma partição uniforme (5.1) e seja P_n o seu polinómio interpolador de Lagrange. Provámos, no exercício anterior, que

$$\max_{x\in[a,b]}|f(x)-P_n(x)| \leq \frac{M_{n+1}}{4(n+1)n^{n+1}}(b-a)^{n+1},$$

para todo o $x \in [a,b]$. Se existir uma constante positiva M tal que

$$M_{n+1} \leq M, \qquad \forall n \in \mathbb{N}, \tag{5.10}$$

concluímos que

$$\lim_{n\to+\infty}\left(\max_{x\in[a,b]}|f(x)-P_n(x)|\right) \leq \lim_{n\to+\infty}\left(\frac{M}{4(n+1)n^{n+1}}(b-a)^{n+1}\right) = 0.$$

Neste caso, o processo de interpolação é convergente, isto é, o aumento do grau do polinómio implica um aumento de precisão.

No entanto existem funções para as quais não podemos concluir que um aumento do grau do provoca um aumento da proximidade do polinómio interpolador com a função interpolada. Isso acontece quando não é possível encontrar um majorante (5.10) para as derivadas da função. O Exercício 5.43 ilustra esta situação.

5.1.3 Fórmula de Newton

Consideremos as seguintes funções

$$\psi_0(x) = 1, \qquad \psi_i(x) = \prod_{j=0}^{i-1}(x-x_i), \quad i = 1,\ldots,n.$$

Atendendo a que o conjunto $\{\psi_i\}_{i=0}^n$ constitui uma base do conjunto dos polinómios de grau inferior ou igual a n (prove), existem constantes $c_i, i = 0,\ldots,n$, tais que o polinómio interpolador de Lagrange é dado por

$$P_n(x) = \sum_{i=0}^{n} c_i\psi_i(x). \tag{5.11}$$

Para determinar c_0 note-se que, se $P_n(x)$ poder ser escrito na forma (5.11), temos que

$$c_0 = P_n(x_0) = f(x_0).$$

De forma similar temos que c_1 pode ser determinado calculando P_n no ponto x_1. Assim

$$f(x_0) + c_1(x_1 - x_0) = P_n(x_1) = f(x_1) \Rightarrow c_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Denotando por $f[x_0, x_1]$ a diferença dividida de primeira ordem

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

e prosseguindo de forma idêntica deduzimos que

$$c_2 = \frac{f(x_2) - f(x_0) - f[x_0, x_1](x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0},$$

que denotamos por $f[x_0, x_1, x_2]$. Podemos deste modo obter um processo recursivo para a determinação dos coeficientes do polinómio se atendermos à seguinte definição.

Definição 5.2 (Diferenças divididas) *Seja f uma função definida nos pontos da partição (5.1) do intervalo $[a, b] \subseteq \mathbb{R}$. A*

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$$

chama-se diferença dividida de primeira ordem *de f relativamente aos argumentos x_i e x_{i+1}. As diferenças divididas de ordem superior definem-se recursivamente. Assim, define-se* diferença dividida de ordem k *relativamente aos argumentos $x_i, x_{i+1}, \ldots x_{i+k}$, com $i+k < n$, por*

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

Usando a definição anterior pode demonstrar-se que

$$c_i = f[x_0, \ldots, x_i], \qquad i = 1, \ldots, n.$$

Substituindo este valor na expressão (5.11) que define $P_n(x)$ obtemos

$$\begin{aligned} P_n(x) &= f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\quad + \cdots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1}) \\ &= f(x_0) + \sum_{i=1}^{n} f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1}(x - x_j), \end{aligned} \tag{5.12}$$

conhecida por fórmula interpoladora de Newton das diferenças divididas. Abusivamente é usual designar por polinómio interpolador de Newton o polinómio calculado por (5.12).

As diferenças divididas são usualmente dadas tabela das diferenças divididas

x_i	$f(x_i)$	$f_{i,i+1}$	$f_{i,i+2}$	$f_{i,i+3}$	$\cdots$
x_0	f_0				
		$f_{0,1}$			
x_1	f_1		$f_{0,2}$		
		$f_{1,2}$		$f_{0,3}$	
x_2	f_2		$f_{1,3}$		$\cdots$
		$f_{2,3}$		$\cdots$	
x_3	f_3		$\cdots$		$\cdots$
		$\cdots$		$f_{n-3,n}$	
$\cdots$	$\cdots$		$f_{n-2,n}$		
		$f_{n-1,n}$			
x_n	f_n				

,

onde $f_{i,j} = f[x_i, x_j]$. Essa tabela pode ser obtida pelo seguinte algoritmo.

Algoritmo 5.2 Diferenças divididas

Dados: n, x_i, $i = 0, 1, \ldots, n$
$f_0 := f(x_0)$
Para i de 1 até n fazer
    $f_i := f(x_i)$
    Para j de $i-1$ até n fazer
        $f_{j,i} := (f_{j+1,i} - f_{j,i-1})/(x_i - x_j)$
Resultado: $f_{j,i}$, $i = 1, \ldots, n$, $j = i-1, \ldots, n$

Notemos que os coeficientes da fórmula de Newton estão ao longo da diagonal da tabela das diferenças divididas.

Um resultado importante respeitante às diferenças divididas é o seguinte.

Teorema 5.3 *As diferenças divididas são invariantes para qualquer permutação dos índices de suporte.*

Demonstração: Com efeito, tem-se que

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f(x_i) - f(x_{i+1})}{x_i - x_{i+1}} = f[x_{i+1}, x_1].$$

Por indução conclui-se facilmente (exercício) que o mesmo acontece para as diferenças divididas de qualquer ordem. □

A demonstração do teorema anterior poderia ter sido feita atendendo ao seguinte exercício que se demonstra por indução.

Exercício 5.7 Seja P_n o polinómio interpolador de $f \in C^{n+1}([a,b])$ de grau inferior ou igual a n nos pontos da partição (5.1) do intervalo $[a,b]$ e w o polinómio nodal dado em (5.6). Mostre que se verifica a igualdade

$$f[x_0, x_1, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{w'(x_i)}.$$

O valor do polinómio interpolador num determinado ponto do seu domínio pode ser dado pelo seguinte algoritmo.

Algoritmo 5.3 Fórmula de Newton das diferenças divididas

Dados: n, $\overline{x}$ e x_i, $i = 0, 1, \ldots, n$
$f_{0,0} := f(x_0)$
Para i de 1 até n fazer
    $f_{i,i} := f(x_i)$
    Para j de $i-1$ até n fazer
        $f_{j,i} := (f_{j+1,i} - f_{j,i-1})/(x_i - x_j)$
$P := f_{0,n}$
Para i de $n-1$ até 0 fazer
    $P := f_{0,i} + (x_i - \overline{x})P$
Resultado: $P(\overline{x}) = P$

Uma grande vantagem do algoritmo de Newton consiste em, uma vez determinado P_n, para determinar P_{n+1} basta fazer

$$P_{n+1}(x) = P_n(x) + f[x_0, x_1, \ldots, x_{n+1}] \prod_{j=0}^{n} (x - x_j).$$

A fórmula (5.12) pode ser calculada de forma mais eficiente se for calculada pelo método de Hörner. De facto, o cálculo do polinómio interpolador usando o fórmula interpoladora de Newton das diferenças divididas na forma encaixada, supondo calculados os coeficientes $f(x_0)$, $f[x_0, x_1]$, ..., $f[x_0, x_1, \ldots, x_n]$, requer apenas $2n$ adições/subtracções e n multiplicações/divisões.

Exercício 5.8 Escreva a fórmula interpoladora de Newton das diferenças divididas usando o método de Hörner.

Exercício 5.9 Dada a tabela

x_i	1	-1	-2
$f(x_i)$	0	-3	-4

, determine uma aproximação para $f(0)$, usando interpolação quadrática.

Resolução: Temos

x_i	$f(x_i)$	$f_{i,i+1}$	$f_{i,i+2}$
1	0		
		$3/2$	
-1	-3		$1/6$
		1	
-2	-4		

.

Assim

$$P_2(x) = 0 + \frac{3}{2}(x-1) + \frac{1}{6}(x^2-1) = (x-1)\left(\frac{3}{2} + \frac{1}{6}(x+1)\right).$$

Temos então que $f(0) \approx P_2(0) = -\frac{5}{3}$.

Apresentámos três processos distintos para a construção do polinómio interpolador de Lagrange de grau n quando são conhecidos $n+1$ valores de uma dada função. Dos processos apresentados aquele que se mostra menos eficiente é o método matricial pois não tem uma forma explícita de determinar os coeficientes do polinómio interpolador. Mais ainda, a determinação destes coeficientes é feita recorrendo à resolução de um sistema de equações lineares em que a matriz deste sistema pode ser mal condicionada.

Vamos agora particularizar o problema ao caso em que os nodos de interpolação estão igualmente distanciados.

Quando os pontos $x_0, x_1, \ldots, x_n$ estão igualmente distanciados, isto é, quando $x_i - x_{i-1} = h$, para $i = 1, \ldots, n$, a fórmula (5.12) pode ser dada em termos dos chamados operadores de diferenças finitas. Dentro da classe desses operadores vamos apenas considerar o operador diferença progressiva.

Definição 5.3 *Seja f uma função definida em $[a, b] \subseteq \mathbb{R}$. O* operador diferença progressiva *define-se por recursão da seguinte forma: a*

$$\Delta f(x) = f(x + h) - f(x)$$

chama-se diferença progressiva de primeira ordem *de f; a* diferença progressiva de ordem k *é definida por*

$$\Delta^k f(x) = \Delta^{k-1}(\Delta f(x)).$$

Exercício 5.10 Prove (por indução) que se f for uma função real definida em $[a, b] \subseteq \mathbb{R}$ e $x_0, x_1, \ldots, x_n$ são pontos de $[a, b]$ igualmente distanciados, com $x_{i-1} - x_i = h$, $i = 1, \ldots, n$, então

$$f[x_0, \ldots, x_k] = \frac{\Delta^k f(x_0)}{k! h^k}, \tag{5.13}$$

para todo o $k \in \{1, \ldots, n\}$.

Substituindo (5.13) em (5.12) temos que

$$\begin{aligned} P_n(x) = \; & f(x_0) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2h^2}(x - x_0)(x - x_1) \\ & + \cdots + \frac{\Delta^n f(x_0)}{n! h^n}(x - x_0)(x - x_1) \cdots (x - x_{n-1}) \\ = \; & f(x_0) + \sum_{i=1}^{n} \frac{\Delta^i f(x_0)}{i! h^i} \prod_{j=0}^{i-1} (x - x_j). \end{aligned} \tag{5.14}$$

Esta fórmula é conhecida por fórmula interpoladora de Newton das diferenças progressivas.

As diferenças progressivas podem ser dadas pela seguinte tabela, conhecida por tabela das diferenças progressivas.

x_i	$f(x_i)$	$\Delta f(x_i)$	$\Delta^2 f(x_i)$	$\Delta^3 f(x_i)$	$\cdots$
x_0	f_0				
		$\Delta f(x_0)$			
x_1	f_1		$\Delta^2 f(x_0)$		
		$\Delta f(x_1)$		$\Delta^3 f(x_0)$	
x_2	f_2		$\Delta^2 f(x_1)$		$\cdots$
		$\Delta f(x_3)$		$\cdots$	
x_3	f_3		$\cdots$		$\cdots$
		$\cdots$		$\Delta^3 f(x_{n-3})$	
$\cdots$	$\cdots$		$\Delta^2 f(x_{n-2})$		
		$\Delta f(x_{n-1})$			
x_n	f_n				

Exercício 5.11 Construa um algoritmo para determinar o valor do polinómio interpolador num determinado ponto do seu domínio usando a fórmula interpoladora de Newton das diferenças progressivas.

Exercício 5.12 Mostre

$$\Delta \arctan x = \arctan \frac{h}{1 + xh + x^2}.$$

Resolução: Vamos provar que

$$\tan\left(\Delta \arctan x\right) = \frac{h}{1 + xh + x^2}.$$

Como

$$\tan\left(\Delta \arctan x\right) = \tan\left(\arctan\left(x + h\right) - \arctan x\right) = \frac{h}{1 + xh + x^2}.$$

5.2 Interpolação de Chebyshev

Uma questão interessante consiste em saber como diminuir os erro de interpolação sem aumentar o número de pontos de suporte. A fórmula (5.8) mostra que o erro de interpolação depende tanto do máximo de $|f^{(n+1)}(x)|$, para todo o x pertencente ao intervalo de interpolação, como de

$$\max_{x \in [a,b]} |w(x)| \tag{5.15}$$

(que depende da escolha dos pontos de interpolação). A questão interessante está em saber, para um dado n, qual a escolha dos pontos de interpolação que minimiza (5.15). A resposta pode ser dada à custa dos chamados polinómios de Chebyshev.

Para $n = 0, 1, 2, \ldots$ e $x \in [-1, 1]$ os polinómios de Chebyshev da grau n são definidos pela relação

$$T_n(x) = \cos\left(n \arccos x\right).$$

Uma forma simples de provar que T_n é, de facto, um polinómio, é atendendo à fórmula de recorrência (ver Exercício 5.13)

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \qquad n = 1, 2, \ldots. \tag{5.16}$$

Exercício 5.13 Obtenha a fórmula de recorrência (5.16) e conclua que T_n é, de facto, um polinómio.

Exercício 5.14 Mostre que o polinómio de Chebyshev T_n tem os seus zeros localizados nos pontos $x_k = \cos\frac{(2k-1)\pi}{2n}$, $k = 1, \ldots, n$, e os extremos localizados em $x'_k = \cos\frac{k\pi}{n}$, $k = 0, \ldots, n$, nos quais $T_n(x'_k) = (-1)^k$.

Da definição de polinómio de Chebyshev resulta imediatamente que $|T_n(x)| \leq 1$, $n = 0, 1, 2, \ldots$. Assim sendo, como $T_n(1) = 1$, temos que $\max_{x\in[-1,1]} |T_n(x)| = 1$. Além disso, atendendo ao Exercício 5.14, os zeros dos polinómios de Chebyshev estão todos no intervalo $[-1, 1]$.

É fácil provar que o coeficiente do termo de maior grau de T_n é $a_n = 2^{n-1}$. Assim sendo, o polinómio $\widetilde{T}_n = 2^{1-n}T_n$ é mónico, isto é, o seu coeficiente do termo de maior grau é igual à unidade. Designemos por $\widetilde{\mathcal{P}}_n([a, b])$ a classe dos polinómios mónicos de grau menor ou igual a n em $[a, b]$.

Teorema 5.4 *O polinómio $\widetilde{T}_n$ é de todos os polinómios de $\widetilde{\mathcal{P}}_n([-1, 1])$ o que tem menor norma, isto é,*

$$\max_{x\in[-1,1]} |\widetilde{T}_n(x)| \leq \max_{x\in[-1,1]} |\widetilde{P}(x)|, \qquad \forall \widetilde{P} \in \widetilde{\mathcal{P}}_n([-1, 1]).$$

Demonstração: Sabemos que $\max_{x\in[-1,1]} |\widetilde{T}_n(x)| = 2^{1-n}$. Suponhamos que existe $\widetilde{P} \in \widetilde{\mathcal{P}}_n([-1, 1])$ tal que $\max_{x\in[-1,1]} |\widetilde{P}(x)| < 2^{1-n}$ e seja $Q = \widetilde{T}_n - \widetilde{P}$. Então o grau de Q é menor ou igual a $n - 1$. Por outro lado, para os valores de x'_k dados no Exercício 5.14,

$$Q(x'_k) = \widetilde{T}_n(x'_k) - \widetilde{P}(x'_k) = (-1)^k 2^{1-n} - \widetilde{P}(x'_k).$$

Assim sendo, o polinómio Q tem n zeros pois tem sinais alternados em n intervalos e é uma função contínua. Logo Q é o polinómio nulo, o que prova o resultado. □

Se considerarmos o intervalo $[a, b]$ em vez do intervalo $[-1, 1]$ há que efectuar a mudança de variável

$$t = \frac{a+b}{2} + \frac{b-a}{2}x.$$

O Teorema 5.4 e o Exercícios 5.14 permitem-nos afirmar, atendendo a que w dado por (5.6) é um polinómio mónico, que (5.15) é mínimo quando se consideram os pontos de suporte

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2i+1)\pi}{2n+2}, \qquad i = 0, \ldots, n.$$

Neste caso o erro é dado por

$$\max_{x\in[a,b]} |f(x) - P_n(x)| \leq \frac{(b-a)^n}{2^{n+1}(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)|,$$

com $P_n(x)$ o polinómio interpolador de Lagrange de f nos pontos dados.

O fenómeno de interpolação também é muito sensível a erros dos dados $y_i = f(x_i)$, $i = 0, \ldots, n$, e a escolha criteriosa dos pontos de suporte pode, também neste aspecto, ser importante. Suponhamos que o cálculo do polinómio interpolador é efectuado com os valores

$$\widehat{y}_i = y_i(1 + \epsilon_i), \qquad |\epsilon_i| < \epsilon.$$

Assim, os polinómios que passam por (x_i, y_i) e $(x_i, \widehat{y}_i)$ são dados, respectivamente, por

$$P_n(x) = \sum_{i=0}^{n} y_i \ell_i(x) \qquad \text{e por} \qquad \widehat{P}_n(x) = \sum_{i=0}^{n} \widehat{y}_i \ell_i(x).$$

Como tal,

$$|\widehat{P}_n(x) - P_n(x)| \leq \epsilon \max_{i=0,\ldots,n} |y_i| \sum_{i=0}^{n} |\ell_i(x)|.$$

Temos então que a função $\sum_{i=0}^{n} |\ell_i(x)|$ descreve o factor de amplificação dos erros dos dados. O seu valor máximo $\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |\ell_i(x)|$ é chamado constante de Lebesgue, em homenagem a Henri Léon Lebesgue (1875-1941), associada aos pontos de interpolação dados e ao intervalo $[a, b]$. Esta constante pode ser calculada numericamente.

Exercício 5.15 Mostre numericamente que, quando se consideram pontos igualmente distanciados no intervalo $[a, b]$, se tem $\Lambda_{20} \simeq 3 \times 10^4$ e $\Lambda_{40} \simeq 10^{10}$ e quando se consideram os pontos de Chebyshev $\Lambda_n \leq 3$ $(n \leq 30)$ e $\Lambda_n \leq 4$ $(n \leq 100)$.

5.3 Interpolação trigonométrica e FFT

Pretende-se aproximar uma função periódica $f : [0, 2\pi] \longrightarrow \mathbb{C}$, com $f(0) = f(2\pi)$, por um polinómio trigonométrico $\widetilde{f}$ que interpola f nos $n + 1$ pontos $x_j = \frac{2\pi j}{n+1}$, $j = 0, \ldots, n$, ou seja, tal que

$$\widetilde{f}(x_j) = f(x_j), \qquad j = 0, \ldots, n.$$

Um polinómio trigonométrico pode escrever-se na forma

$$\widetilde{f}(x) = \begin{cases} \displaystyle\sum_{k=-M}^{M} c_k e^{ikx}, & M = \frac{n}{2}, n \text{ par} \\ \displaystyle\sum_{k=-M-1}^{M+1} c_k e^{ikx}, & M = \frac{n-1}{2}, n \text{ ímpar} \end{cases},$$

onde os coeficientes c_k são números complexos e $c_{M+1} = c_{-M-1}$. De forma mais compacta, um polinómio trigonométrico representa-se por

$$\widetilde{f}(x) = \sum_{k=-M-\mu}^{M+\mu} c_k e^{ikx}, \tag{5.17}$$

com $\mu = 0$, se n é par, ou $\mu = 1$, se n é ímpar. Devido à sua analogia com as séries de Fourier, estudadas por Jean Baptiste Joseph Fourier (1768-1830), a $\widetilde{f}$ chama-se série de Fourier discreta.

Note-se que, se f for uma função real, os coeficientes c_k são tais que c_{-k} é o conjudado de c_k, com $k = 1, ..., M + \mu$.

Considerando a condição de interpolação em $x_j = jh$, com $h = \frac{2\pi}{n+1}$, tem-se que

$$\widetilde{f}(x_j) = \sum_{k=-M-\mu}^{M+\mu} c_k e^{ikjh} = f(x_j), \qquad j = 0, \ldots, n. \tag{5.18}$$

Vamos mostrar que os coeficentes da série de Fourier discreta (5.17) verificam

$$c_k = \frac{1}{n+1}\sum_{j=0}^{n} f(x_j)e^{-ikjh}, \qquad k = -M-\mu, \ldots, M+\mu.$$

Multiplicando por $e^{-imx_j} = e^{-imjh}$, com m um inteiro entre 0 e n, e tomando somatórios, em j, em ambos os membros de (5.18) tem-se que

$$\sum_{j=0}^{n} f(x_j)e^{-imjh} = \sum_{j=0}^{n}\sum_{k=-M-\mu}^{M+\mu} c_k e^{i(k-m)jh} = \sum_{k=-M-\mu}^{M+\mu} c_k \sum_{j=0}^{n} e^{i(k-m)jh}.$$

O resultado pretendido resulta do facto de

$$\sum_{j=0}^{n} e^{i(k-m)jh} = (n+1)\delta_{km},$$

onde δ_{km} é o símbolo de Kronecker. Para provar este facto, notemos que, se $k = m$,

$$\sum_{j=0}^{n} e^{i(k-m)jh} = \sum_{j=0}^{n} 1 = n+1.$$

Se $k \neq m$,

$$\sum_{j=0}^{n} e^{i(k-m)jh} = \frac{1-e^{i(k-m)h(n+1)}}{1-e^{i(k-m)h}} = \frac{1-e^{i(k-m)2\pi}}{1-e^{i(k-m)h}} = 0.$$

Ao conjunto dos coeficientes $\{c_k\}$ da série de Fourier discreta (5.17) chamamos transformada discreta de Fourier. A transformada discreta de Fourier pode ser calculada com um número de operações da ordem $n\log_2 n$ usando um algoritmo designado por FFT (de *Fast Fourier Transform*). A transformada rápida de Fourier não é mais do que a transformada discreta de Fourier calculada pelo algoritmo FFT. O MatLab possui o comando `fft` onde está implementado esse algoritmo. O algoritmo mais famoso para obter a FFT é de 1965 e é devido a James Cooley (1926-) e John Wilder Tukey (1915-2000).

A transformação de Fourier inversa, pela qual os valores $\{f(x_k)\}$ são obtidos à custa dos valores $\{c_k\}$, é definida, em MatLab, pela função `ifft`.

Um cuidado a ter com o comando `fft` é o seginte. Se considerarmos

$$F = [f(x_0), f(x_1), \ldots, f(x_n)]^T,$$

ao executar `TF = fft(F)`, obtém-se

$$TF = (n+1)[c_0, \cdots, c_{M+\mu}, c_{-M}, \ldots, c_{-1}]^T.$$

Se pretendermos obter os ceficientes $\{c_k\}$ pela ordem com que aparecem na série de Fourier discreta (5.17) temos que executar o comando `TF = 1/(n + 1) * fftshift(fft(F))`.

Em muitos casos, a precisão da interpolação trigonométrica pode degradar-se muito devido ao fenómeno conhecido com aliasing. Este fenómeno pode ocorrer quando a função a aproximar for a soma de várias componentes com frequências distintas. S o número de nós não for suficiente para resolver as frequências mais altas, estas podem intreferir com as baixas frequências, o que dá origem a interpolações imprecisas. Nesse caso, é preciso aumentar o número de nós de interpolação.

Exercício 5.16 Calcule a transformada de Fourier discreta de $y = [-1, -1, 1, 1]^T$.

5.4 Interpolação seccionalmente linear

Consideremos um intervalo $[a, b]$ e uma partição dada por (5.1). Designemos por polinómio segmentado linear (ou função linear por segmentos) na partição (5.1), uma função contínua em $[a, b]$ que, quando restringida a cada um dos intervalos $[x_i, x_{i+1}]$, $i = 0, ..., n-1$, da partição, coincide com um polinómio de grau menor ou igual a um (polinómio que, em geral, varia com i).

Consideremos agora o problema da interpolação. Seja f uma função conhecida nos pontos da partição (5.1). Pelo que foi visto na secção anterior, é óbvio que existe um e um só polinómio segmentado linear P_1^h tal que

$$P_1^h(x_i) = f(x_i), \qquad i = 0, 1, \ldots, n.$$

Nestas condições, P_1^h é chamado o polinómio interpolador (de Lagrange) segmentado linear de f nos pontos de (5.1).

Temos que

$$P_1^h(x) = \begin{cases} P_{10}^h(x), & x \in [x_0, x_1] \\ P_{11}^h(x), & x \in [x_1, x_2] \\ \vdots & \vdots \\ P_{1i}^h(x), & x \in [x_i, x_{i+1}] \\ \vdots & \vdots \\ P_{1n-1}^h(x), & x \in [x_{n-1}, x_n] \end{cases},$$

onde P_{1i}^h pode ser escrito na forma seguinte

$$P_{1i}^h(x) = f(x_i) + f[x_i, x_{i+1}](x - x_i),$$

ou ainda (fórmula de Lagrange)

$$P_{1i}^h(x) = f(x_i)\frac{x - x_{i+1}}{x_i - x_{i+1}} + f(x_{i+1})\frac{x - x_i}{x_{i+1} - x_i}.$$

O que podemos dizer quanto ao erro que se comete ao aproximar f pelo seu polinómio interpolador segmentado linear?

Suponhamos que $x \in [x_i, x_{i+1}]$. Temos então que, nesse intevalo,

$$\max_{x \in [x_i, x_{i+1}]} \left| f(x) - P_{1i}^h(x) \right| \leq \frac{M_2^{(i)}}{2} \max_{x \in [x_i, x_{i+1}]} |(x - x_i)(x - x_{i+1})|$$

com

$$M_2^{(i)} = \max_{x \in [x_i, x_{i+1}]} \left| f^{(2)}(x) \right|, \qquad i = 0, \ldots, n-1.$$

Mas, como vimos,

$$\max_{x \in [x_i, x_{i+1}]} |(x - x_i)(x - x_{i+1})| = \frac{1}{4}(x_{i+1} - x_i)^2$$

e, com tal

$$\max_{x \in [x_i, x_{i+1}]} \left| f(x) - P_{1i}^h(x) \right| \leq \frac{M_2^{(i)}}{8} h_i^2,$$

com $h_i = x_{i+1} - x_i$, $i = 0, \ldots, n-1$.

Consideremos agora $x \in [a,b]$. Atendendo ao que foi dito,conclui-se imediatamente que

$$\max_{x\in[a,b]} \left| f(x) - P_1^h(x) \right| \leq \frac{M_2}{8} h^2,$$

onde $M_2 = \max_{i=0,\ldots,n-1} M_2^{(i)}$ e $h = \max_{i=0,\ldots,n-1} h_i$.

Este limite superior para o erro permite demonstrar que o processo de interpolação linear por segmentos é convergente. De facto, se $f^{(2)}$ é limitada, à medida que o número de pontos da partição aumenta (h diminui) o erro tende para zero, ou seja, o polinómio segmentado linear tende para a função a interpolar uniformemente em $[a,b]$.

A interpolação linear segmentada possui vantagens em relação à interpolação (global) de Lagrange. Note-se que, se n é muito grande o cálculo do polinómio interpolador de Lagrange (global) P_n envolve muito mais operações que o cálculo do polinómio interpolador linear segmentado S. Além disso, como foi visto, o facto de n aumentar não implica que o polinómio interpolador de Lagrange P_n tenda para a função a interpolar, mesmo que essa função seja infinitamente diferenciável. A desvantagem que o processo da interpolação segmentada linear apresenta relativamente à interpolação de Lagrange é que o polinómio P_n é infinitamente diferenciável enquanto que P_n^h pode não ter (e, em geral, não tem) derivadas contínuas nos pontos da partição.

5.5 Interpolação de Hermite

O objectivo da interpolação de Hermite, chamada assim em honra de Charles Hermite (1822-1901), é o de representar uma função f por um polinómio que seja interpolador de f em alguns pontos do seu domínio e que a sua derivada seja interpolador da derivada de f nesses mesmos pontos. Isto é, supondo que f é diferenciável, vamos procurar um polinómio H tal que

$$\begin{cases} f(x_i) &= H(x_i) \\ f'(x_i) &= H'(x_i) \end{cases}, \qquad i = 0,1,\ldots,n. \tag{5.19}$$

Quando tal situação acontece dizemos que f e H são funções que 2-osculam (osculam 2 vezes) os pontos x_i, $i = 0,1,\ldots,n$, ou que H é um polinómio 2-osculador de f nos pontos x_i, $i = 0,1,\ldots,n$.

O próximo teorema estabelece a existência e unicidade do polinómio de grau inferior ou igual a $2n+1$ que verifica (5.19). Além disso, indica-nos um processo que permite a sua determinação.

Teorema 5.5 *Seja $f \in C^{2n+2}([a,b])$ e $x_0, x_1, \ldots, x_n$ pontos distintos em $[a,b]$. Existe um e um só polinómio H_{2n+1} de grau menor ou igual a $2n+1$ que verifica (5.19).*

Demonstração: Atendendo às condições impostas, o polinómio terá que ser de grau inferior ou igual a $2n+1$. Para provar a sua existência vamos considerar as funções

$$h_{1i}(x) = [1 - 2\ell_i'(x_i)(x - x_i)]\ell_i(x)^2 \qquad \text{e} \qquad h_{2i}(x) = (x - x_i)\ell_i(x)^2, \qquad i = 0,\ldots,n,$$

com ℓ_i, $i = 0,\ldots,n$, os polinómios de Lagrange (5.3). Como se pode verificar facilmente

$$h_{1i}(x_j) = \delta_{i,j}, \quad h_{1i}'(x_j) = 0, \qquad i,j = 0,\ldots,n,$$

e

$$h_{2i}(x_j) = 0, \quad h_{2i}'(x_j) = \delta_{i,j}, \qquad i,j = 0,\ldots,n.$$

Assim, o polinómio

$$H_{2n+1}(x) = \sum_{i=0}^{n} [f(x_i)h_{1i}(x) + f'(x_i)h_{2i}(x)]$$

tem grau inferior ou igual a $2n+1$ e verifica (5.19).

Falta apenas provar a unicidade. Seja Q_{2n+1} outro polinómio de grau inferior ou igual a $2n+1$ que verifica (5.19) e

$$R_{2n+1}(x) = H_{2n+1}(x) - Q_{2n+1}(x).$$

Como $R_{2n+1}(x_i) = R'_{2n+1}(x_i) = 0$, para $i = 0, \ldots, n$, temos que este polinómio de grau inferior ou igual a $2n+1$ tem $2n+2$ zeros o que implica que terá que ser o polinómio nulo. Assim sendo, provámos a unicidade pretendida. □

O único polinómio de grau menor ou igual a $2n+1$ que verifica as condições (5.19) é também chamado polinómio interpolador de Hermite de f nos pontos $x_0, x_1, \ldots, x_n$.

Note-se que, tal como na interpolação de Lagrange, se m for o número de condições impostas para a determinação do polinómio interpolador, o seu grau é $m-1$.

Exercício 5.17 Mostre que o polinómio de Hermite de grau mínimo (n=1) é dado por

$$H_3(x) = f(x_0)h_{10}(x) + f(x_1)h_{11}(x) + f'(x_0)h_{20}(x) + f'(x_1)h_{21}(x),$$

com

$$h_{10}(x) = \left(1 - 2\frac{x - x_0}{x_0 - x_1}\right)\frac{(x-x_1)^2}{(x_0-x_1)^2},$$

$$h_{11}(x) = \left(1 - 2\frac{x - x_1}{x_1 - x_0}\right)\frac{(x-x_0)^2}{(x_1-x_0)^2},$$

$$h_{20}(x) = (x - x_0)\frac{(x-x_1)^2}{(x_0-x_1)^2},$$

$$h_{21}(x) = (x - x_1)\frac{(x-x_0)^2}{(x_1-x_0)^2}.$$

O estudo do erro na interpolação de Hermite consiste na generalização do estudo efectuado para a interpolação de Lagrange de acordo com o seguinte teorema.

Teorema 5.6 *Seja H_{2n+1} o polinómio, de grau menor ou igual a $2n+1$ interpolador de Hermite da função f nos pontos distintos $x_0, x_1, \ldots, x_n \in [a,b]$. Se $f \in C^{2n+2}([a,b])$ então para cada $x \in [a,b]$ existe $\xi = \xi(x) \in]a,b[$ tal que*

$$e(x) = f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}w^2(x), \tag{5.20}$$

onde w é a função dada por (5.6).

Tal como no caso da interpolação de Lagrange, pelo teorema anterior podemos determinar um majorante para o erro cometido ao substituir f pelo seu polinómio interpolador de Hermite de grau n, H_{2n+1}. De facto, de (5.20) sai que:

$$\max_{x\in[a,b]} |f(x) - H_{2n+1}(x)| \leq \frac{M_{2n+2}}{(2n+2)!} \max_{x\in[a,b]} |w^2(x)|,$$

onde

$$M_{2n+2} = max_{x\in[a,b]}|f^{(2n+2)}(x)|.$$

Se o objectivo for determinar o erro apenas num ponto $\overline{x} \in [a,b]$, então

$$|f(\overline{x}) - H_{2n+1}(\overline{x})| \leq \frac{M_{2n+2}}{(2n+2)!}|w^2(\overline{x})|.$$

Atendendo a que $x, x_j \in [a,b]$, temos que $|x - x_j| \leq (b-a)$ e portanto

$$\max_{x\in[a,b]} |f(x) - H_{2n+1}(x)| \leq \frac{M_{2n+2}}{(2n+2)!}(b-a)^{2n+2}.$$

Podemos também concluir, uma vez que

$$\max_{x\in[a,b]} |w(x)| \leq \frac{h^{n+1}n!}{4},$$

com $h = \max\limits_{i=1,\ldots,n}(x_i - x_{i-1})$, que

$$\max_{x\in[a,b]} |f(x) - H_{2n+1}(x)| \leq M_{2n+2}\frac{h^{2n+2}(n!)^2}{16(2n+2)!}.$$

Observamos que dependendo do comportamento de M_{2n+2} podemos, ou não, concluir que o aumento do grau do polinómio interpolador de Hermite implica uma diminuição do erro cometido ao aproximar a função por este polinómio. Uma forma de minimizar o erro consiste na utilização de polinómios interpoladores segmentados.

5.5.1 Interpolação segmentada de Hermite

Consideremos um intervalo $[a,b]$ e uma partição dada por (5.1). Designemos por polinómio segmentado cúbico (ou função cúbica por segmentos) na partição (5.1), uma função contínua em $[a,b]$ que, quando restringida a cada um dos intervalos $[x_i, x_{i+1}]$, $i = 0, ..., n-1$, da partição, coincide com um polinómio de grau menor ou igual a três.

Seja f uma função conhecida nos pontos da partição (5.1). Como se sabe, existe um e um só polinómio segmentado cúbico H_3^h tal que

$$\left\{ \begin{array}{lcl} H_3^h(x_i) & = & f(x_i) \\ (H_3^h)'(x_i) & = & f'(x_i) \end{array} \right. , \qquad i = 0, 1, \ldots, n.$$

Nestas condições, H_3^h é chamado o polinómio interpolador (de Hermite) segmentado cúbico de f nos pontos de (5.1).

Temos que

$$H_3^h(x) = \left\{ \begin{array}{ll} H_{30}^h(x), & x \in [x_0, x_1] \\ H_{31}^h(x), & x \in [x_1, x_2] \\ \vdots & \vdots \\ H_{3i}^h(x), & x \in [x_i, x_{i+1}] \\ \vdots & \vdots \\ H_{3n-1}^h(x), & x \in [x_{n-1}, x_n] \end{array} \right. ,$$

onde H_{3i}^h pode ser escrito na forma seguinte

$$H_{3i}^h(x) = f(x_i)h_{1i}(x) + f(x_{i+1})h_{1i+1}(x) + f'(x_i)h_{2i}(x) + f'(x_{i+1})h_{2i+1}(x).$$

Exercício 5.18 Mostre que o erro que se comete ao aproximar $f \in C^4([a,b])$ pelo seu polinómio interpolador segmentado de Hermite cúbico na partição (5.1) é dado por

$$\max_{x\in[a,b]} \left| f(x) - H_3^h(x) \right| \leq \frac{M_4}{384} h^4,$$

onde $M_4 = \max_{x\in[a,b]} \left| f^{(4)}(x) \right|$ e $h = \max_{i=1,\ldots,n} (x_i - x_{i-1})$.

5.5.2 Polinómio interpolador de Hermite e diferenças divididas

A obtenção do polinómio interpolador de Hermite pode ser feita de várias maneiras. Vamos apresenta-la neste curso numa forma que generaliza o polinómio interpolador de Newton das diferenças divididas.

Consideremos a mudança de variável $z_0 = x_0$, $z_1 = x_0$, $z_2 = x_1$, $z_3 = x_1$, ..., $z_{2n} = x_n$, $z_{2n+1} = x_n$. Uma vez que

$$z_{2i} = z_{2i+1} = x_i, \qquad i = 0, \ldots, n,$$

não podemos definir as diferenças divididas

$$f[z_{2i}, z_{2i+1}] = f[x_i, x_i].$$

No entanto, atendendo a que

$$\lim_{x\to x_i} f[x, x_i] = \lim_{x\to x_i} \frac{f(x) - f(x_i)}{x - x_i} = f'(x_i),$$

podemos definir as diferenças divididas generalizadas para pontos não distintos na forma

$$f[x_i, x_i] = f'(x_i).$$

Pelo Teorema do Valor Médio de Lagrange generalizado podemos ainda definir

$$f[\underbrace{x_i, x_i, \ldots, x_i}_{r+1 \text{ vezes}}] = \frac{f^{(r)}(x_i)}{r!}. \tag{5.21}$$

Com esta notação pode verificar-se facilmente que o polinómio interpolador de Hermite de grau $2n+1$ nos pontos da partição (5.1) é dado por (verifique para $n = 1$)

$$\begin{aligned} H_{2n+1}(x) = \; & f(z_0) + \sum_{i=1}^{2n+1} f[z_0, z_1, \ldots, z_i] \prod_{j=0}^{i-1} (x - z_j) \\ = \; & f(x_0) + f'(x_0)(x - x_0) \\ & + f[x_0, x_0, x_1](x - x_0)^2 + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1) \\ & + \cdots + f[x_0, x_0, \ldots, x_n, x_n](x - x_0)^2(x - x_1)^2 \cdots (x - x_{n-1})^2(x - x_n). \end{aligned}$$

Exercício 5.19 Prove a afirmação anterior para o caso em que se consideram apenas dois pontos de interpolação ($n = 1$).

O polinómio interpolador de Hermite pode assim ser determinado recorrendo à tabela das diferenças divididas generalizadas, tabela essa onde cada ponto aparece repetido duas vezes.

Exercício 5.20 Construa um algoritmo para determinar o valor do polinómio interpolador de Hermite num determinado ponto do seu domínio.

Exercício 5.21 Determine o polinómio interpolador de Hermite de grau mínimo para a função $f(x) = \sin x$ em $[0, \frac{\pi}{2}]$.

Resolução: Temos

x_i	$f(x_i)$	$f[\cdot,\cdot]$	$f[\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\cdot,\cdot]$
0	0			
		1		
0	0		$\frac{4-2\pi}{\pi^2}$	
		$\frac{2}{\pi}$		$\frac{4\pi-16}{\pi^3}$
$\frac{\pi}{2}$	1		$-\frac{4}{\pi^2}$	
		0		
$\frac{\pi}{2}$	1			

.

Logo

$$\begin{aligned} H_3(x) &= x + \frac{4-2\pi}{\pi^2}x^2 + \frac{4\pi-16}{\pi^3}x^2\left(x-\frac{\pi}{2}\right) \\ &= x\left(1 + x\left(-0{,}2313 - 0{,}1107\left(x-\frac{\pi}{2}\right)\right)\right). \end{aligned}$$

5.6 Aproximação por funções *spline* cúbicas

O termo inglês *spline* pode ser traduzido pelo vocábulo "virote". Um virote é um instrumento usado pelos desenhadores para unir um conjunto de pontos do plano.

Seja f uma função definida num intervalo $[a, b]$ onde consideramos a partição (5.1). Matematicamente, o problema de unir pontos do plano com um virote pode ser traduzido da seguinte forma: determinar a função $S : [a, b] \longrightarrow \mathbb{R}$, com $a = x_0$, $b = x_n$, que satisfaz:

[S1] $S(x_i) = f(x_i), \qquad i = 0, \ldots, n;$

[S2] $S \in C^2([a, b])$;

[S3] o princípio de Pierre-Louis Moreau de Maupertuis (1698-1759) da energia mínima, i.e.,

$$\int_a^b (S''(x))^2 dx \leq \int_a^b (R''(x))^2 dx,$$

para toda a função R que satisfaz [S1] e [S2].

Atente-se ao seguinte resultado.

Teorema 5.7 *Sejam $S, R : [a, b] \longrightarrow \mathbb{R}$ duas funções que verificam [S1] e [S2]. Suponhamos que*

$$S''(b)\left(R'(b) - S'(b)\right) = S''(a)\left(R'(a) - S'(a)\right)$$

e que S é um polinómio de grau 3 em cada sub-intervalo da partição dada. Então temos que

$$\int_a^b (S''(x))^2 dx \leq \int_a^b (R''(x))^2 dx.$$

Demonstração: Temos que

$$\int_a^b (g''(x))^2 dx - \int_a^b (S''(x))^2 dx = \int_a^b (g''(x) - S''(x))^2 dx + 2\int_a^b S''(x)(g''(x) - S''(x))dx.$$

Integrando por partes último integral do segundo membro vem

$$\begin{aligned}\int_a^b S''(x)(g''(x) - S''(x))dx &= S''(b)[g'(b) - S'(b)] - S''(a)[g'(a) - S'(a)] \\ &\quad - \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} S'''(x)(g''(x) - S''(x))dx.\end{aligned}$$

Ora, atendendo às hipóteses do teorema,

$$\int_a^b S''(x)(g''(x) - S''(x))dx = 0$$

e, como tal,

$$\int_a^b (S''(x))^2 dx = \int_a^b (g''(x))^2 dx - \int_a^b (g''(x) - S''(x))^2 dx,$$

o que permite concluir o pretendido. □

Este teorema mostra que os candidatos à resolução de [S1]–[S3] são as funções pertencentes a $C^2([a,b])$ que são polinómios de grau 3 em cada intervalo da partição. Essas funções serão designadas por funções *spline* cúbicas.

Definição 5.4 (Spline) *Uma* função *spline de grau m é um polinómio segmentado de grau m continuamente derivável até à ordem m − 1. Por outras palavras, dada uma partição (5.1), S é uma função spline de grau m se S é um polinómio $S^{(i)}$ de grau m em cada intervalo da partição $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$, e*

$$\frac{d^k}{dx^k} S^{(i+1)}(x) = \frac{d^k}{dx^k} S^{(i)}(x), \qquad k = 0, \ldots, m-1, \quad i = 0, \ldots, n-1.$$

As funções *spline* mais populares são as cúbicas ($m = 3$). Pelas razões apresentadas, serão essas que iremos considerar.

Note-se que, em cada intervalo $[x_i, x_{i+1}]$ a função *spline* cúbica S que interpola f nos pontos da partição (5.1) é um polinómio de grau 3 e, como tal, é definido à custa de 4 parâmetros. Assim, para determinar S de forma única temos que especificar $4n$ parâmteros. Para isso teremos que definir $4n$ equações. Atendendo à definição de função *spline* temos impostas as seguintes equações: $n+1$ equações de interpolação; $n-1$ equações de ligação de S; $n-1$ equações de ligação de S' e $n-1$ equações de ligação de S''. No total temos assim $4n-2$ equações. Para determinar S temos que considerar mais duas condições suplementares. As formas mais usuais de definir essas condições são as seguintes: $S'(a) = f'(a)$ e $S'(b) = f'(b)$ (*spline* completa); $S''(a) = 0$ e $S''(b) = 0$ (*spline* natural).

O seguinte teorema, que apresentamos sem demonstração, establece a existência e unicidade da função *spline* cúbica interpoladora.

Teorema 5.8 *Seja f uma função definida em $[a,b]$. A função* spline *cúbica completa (natural) interpoladora de f nos pontos da partição (5.1) existe e é única.*

O *spline* cúbico interpolador de uma função coincide com a função nos pontos da partição e a sua derivada coíncide com a derivada da função nos extremos do intervalo de partição. No resultado seguinte, estabelecido sem demonstração, é apresentado o comportamento do erro que se comete ao aproximar uma função pelo seu *spline* cúbico interpolador.

Teorema 5.9 *Seja f uma função definida em $[a,b]$ onde considerado $n+1$ pontos igualmente distanciados $x_i, i=0,\ldots,n$, em que $x_i = x_{i-1}+h, i=0,\ldots,n, x_0 = 0$, $h=(b-a)/n$. O erro cometido ao aproximar f pelo seu* spline *cúbico interpolador verifica a*

$$\max_{x\in[a,b]} |f(x) - S(x)| \leq \frac{5}{384} \max_{x\in[a,b]} |f^{(4)}(x)| h^4.$$

Exercício 5.22 Pretende-se interpolar a função f definida por $f(x) = \ln x$, com $x \in [2,2{,}5]$, por um *spline* cúbico completo S numa malha uniforme.

1. Calcule o número mínimo de pontos a usar para garantir que

$$\max_{x\in[2,2,5]} |f(x) - S(x)| \leq 0{,}5 \times 10^{-4}.$$

2. Determine uma aproximação para $f(2{,}3)$ usando o *spline* cúbico completo interpolador de f nos pontos obtidos na alínea anterior.

5.7 Problemas

5.7.1 Exercícios para resolver nas aulas

Exercício 5.23 Considere os seguintes pontos de $\mathbb{R}^2$, $(-3,1)$, $(-2,2)$, $(1,-1)$ e $(3,10)$. Determine o polinómio interpolador de Lagrange $P(x)$ que passa por esses pontos e calcule $P(0)$.

Exercício 5.24 Os dados da tabela seguinte dizem respeito à esperança média de vida na Europa ocidental

Ano	1975	1980	1985	1990
EMV	72,8	74,2	75,2	76,4

.

1. Calcule o polinómio interpolador de Lagrange usando todos os dados da tabela anterior.

2. Usando a alínea anterior, determine uma aproximação para a esperança média de vida em 1982.

3. Mostre que dados $(x_i, f(x_i)), i=0,\ldots,n$ pontos distintos, o polinómio interpolador de Lagrange de grau n, P_n, é o único polinómio de grau inferior ou igual a n que interpola f.

Exercício 5.25 Considere a função $f : [0,\pi] \longrightarrow \mathbb{R}$ definida por $f(x) = \cos(x),\ x \in [0,\pi]$. Determine o número de pontos a considerar no intervalo dado para que o erro absoluto máximo da aproximação de $f(x)$ por um polinómio interpolador nesses pontos seja inferior a $0,05$.

Exercício 5.26 Determine aproximações de $\cos(\frac{\pi}{8})$ usando polinómios interpoladores de Lagrange de grau 2 e 4, no intervalo $[0,\pi]$. Compare os resultados obtidos e indique majorantes para o erro absoluto.

Exercício 5.27 (Matlab) Considere os seguintes pontos de $\mathbb{R}^2$, (-2,1), (-1,0), (1,-3) e (4,8). Determine o polinómio interpolador $P(x)$ que passa por esses pontos e calcule $P(0)$.

Exercício 5.28 (Matlab) Considere a função $f(x) = \sin x$, $x \in [-\pi/2, \pi/2]$. Trace os gráficos dos polinómios interpoladores de f para diferentes valores de n (grau do polinómio).

Exercício 5.29 Seja $\{x_0, x_1, \ldots, x_n\}$ um conjunto de $n+1$ números reais. Mostre que, $\sum_{i=0}^{n} \ell_i(x) = 1$, onde $\ell_i(x)$, $i = 1, \ldots, n$, são os polinómios de Lagrange (5.4).

Exercício 5.30 Seja $\{x_0, x_1, \ldots, x_n\}$ um conjunto de $n+1$ números reais igualmente espaçados. Mostre que

$$\prod_{i=0}^{n} |x - x_i| \leq \frac{n!\ h^{n+1}}{4},$$

sendo h o espaçamento entre aqueles pontos.

Exercício 5.31 Seja $P(x)$ um polinómio de grau inferior ou igual a n e

$$w(x) = (x - x_0)(x - x_1) \cdots (x - x_n),$$

onde x_i, $i = 0, 1, \ldots, n$, são $n+1$ pontos distintos. Mostre que o coeficiente do termo de maior grau de $P(x)$ é dado por

$$a_0 = \sum_{i=0}^{n} \frac{P(x_i)}{w'(x_i)}.$$

Exercício 5.32 Considere a função $f(x) = \sin x$, definida em $[0, \frac{\pi}{2}]$.

1. Determine o menor número de pontos que deve considerar no intervalo dado para que o erro da aproximação de $f(x)$ por um polinómio interpolador nesses pontos seja inferior a 0,1.

2. Sabendo que $\sin(\frac{\pi}{3}) = \frac{\sqrt{3}}{2}$, determine uma aproximação para $\sqrt{3}$ utilizando um polinómio interpolador de ordem 2.

Exercício 5.33 Considere a função $f(x) = \cos(x) + \sin(x)$ e os pontos

$$x_k = -\frac{\pi}{2} + k\frac{\pi}{4}, \quad k = 0, 1, \ldots, 4.$$

Determine um majorante do erro que se comete na aproximação de f por um polinómio interpolador de Lagrange definido nesses pontos.

Exercício 5.34 Seja x_{k-1}, x_k e x_{k+1} três pontos igualmente espaçados, com distância $h/2$, onde são conhecidos os valores de uma função f. Mostre que o polinómio interpolador de Lagrange de grau 2 é dado por

$$\frac{2(x - x_k)(x - x_{k+1})}{h^2} f(x_{k-1}) + \frac{4(x_{k-1} - x)(x - x_{k+1})}{h^2} f(x_k) + \frac{2(x - x_k)(x - x_{k-1})}{h^2} f(x_{k+1}).$$

Exercício 5.35 (Matlab) Um pára-quedista efectuou 5 saltos de diferentes alturas, tendo medido a distância a um alvo constituído por uma circunferência de raio 5 metros traçada no solo. Supondo que as respectivas altura e distância de cada salto satisfazem a seguinte tabela

Altura do salto (m)	1500	1250	1000	750	500
Distância do alvo (m)	35	25	15	10	7

,

recorra à interpolação para estimar a distância do alvo a que o pára-quedista cairia se saltasse de uma altura de 850 m.

Exercício 5.36 (Matlab) Os jactos de água dos repuxos da Avenida Sá da Bandeira descrevem uma trajectória parabólica. Para obter a expressão dessa trajectória foram realizadas as seguintes medições:

Distância (eixo horizontal)	0	1/4	1/3	1	3/2	2
Altura da água	0	21/16	5/3	3	9/4	0

.

Recorra à interpolação para obter a respectiva trajectória.

Exercício 5.37 (Matlab) A temperatura do ar próximo do solo depende da concentração K em ácido carbónico (H_2CO_3). A tabela representa a variação $\delta_K = \theta_K - \theta_{\overline{K}}$ da temperatura média relativamente a uma temperatura de referência $\overline{K}$, para diferentes latitudes e valores de K:

	δ_K			
Latitude	$K = 0{,}67$	$K = 1{,}5$	$k = 2{,}0$	$K = 3{,}0$
65	$-3{,}1$	3,52	6,05	9,3
55	$-3{,}22$	3,62	6,02	9,3
45	$-3{,}3$	3,65	5,92	9,17
35	$-3{,}32$	3,52	5,7	8,82
25	$-3{,}17$	3,47	5,3	8,1
15	$-3{,}07$	3,25	5,02	7,52
5	$-3{,}02$	3,15	4,95	7,3
-5	$-3{,}02$	3,15	4,97	7,35
-15	$-3{,}12$	3,2	5,07	7,62
-25	$-3{,}2$	3,27	5,35	8,22
-35	$-3{,}35$	3,52	5,62	8,8
-45	$-3{,}37$	3,7	5,95	9,25
-55	$-3{,}25$	3,7	6,1	9,5

.

1. Recorra ao polinómio interpolador dos dados para comparar a variação da temperatura, em função da latitude, para os diferentes valores de K. Trace os respectivos gráficos.

2. Use a alínea anterior para estimar o valor da variação de temperatura δ_K para um local cuja latitude é igual a 23. Considere $K = 0{,}67$.

Exercício 5.38 Uma empresa apresenta os seguintes lucros em função das vendas:

Nº peças vendidas (milhares)	1	2	3	4	5
Lucro (milhares de euros)	11,2	15,3	17,1	16,9	15,0

.

Sabendo que o lucro previsto era de 13 mil euros, indique uma aproximação do número de peças que foi necessário vender para atingir esse lucro.

Exercício 5.39 Obtenha um valor aproximado para a raiz de uma função contínua $f(x)$ da qual se conhece apenas os valores apresentados na tabela seguinte:

x_i	-2	0	1
$f(x_i)$	-12,5	1,5	-1

.

Exercício 5.40 Calcule um polinómio interpolador trigonométrico para a extensão periódica da função $f : [0, 2\pi] \longrightarrow \mathbb{R}$ definida por

$$f(x) = \begin{cases} 1, & x \in [\pi, \frac{3\pi}{2}], \\ -1, & \text{caso contrário,} \end{cases}$$

nos pontos $x_j = \frac{2\pi j}{4},\ j = 0, 1, 2, 3.$

Exercício 5.41 (Matlab) Dada a função

$$f(x) = x(x - 2\pi)e^{-x}, \qquad x \in [0, 2\pi],$$

determine a interpoladora trigonométrica de f em 10 nós equidistantes. Compare o respectivo gráfico com o gráfico de f.

Exercício 5.42 (Matlab) Considere a função

$$f(x) = 1/(1 + x^2)$$

definida no intervalo [-5,5].

1. Trace o gráfico de alguns polinómios interpoladores de $f(x)$ em pontos equidistantes e compare-os com o gráfico da função.

2. Repita o procedimento da alínea anterior usando os nós de Chebyshev. Que pode concluir?

Exercício 5.43 (Matlab) Considere a função (de Runge)

$$f(x) = 1/(1 + 25x^2), \qquad x \in [-1, 1].$$

Verifique graficamente que

$$\max_{x\in[-1,1]} |f(x) - P_3(x)| \leq \max_{x\in[-1,1]} |f(x) - P_8(x)|\,,$$

em que P_3 e P_8 são, respectivamente, os polinómios de Lagrange de grau 3 e 8 interpoladores de f em partições uniformes de $[-1, 1]$.

Exercício 5.44 A lei de Ohm diz que a tensão V nas extremidades de uma resistência percorrida por uma corrente eléctrica com intensidade I é directamente proporcional a essa intensidade de corrente. Isso só é verdade para resistências ideais; as resistências reais apresentam comportamentos menos lineares. Na tabela seguinte apresentam-se os resultados para uma resistência concreta:

I	$-1{,}0$	$-0{,}5$	0	$0{,}5$	$1{,}0$
V	-193	-41	0	41	193

.

Atendendo ao comportamento simétrico relativamente à origem, determine o polinómio segmentado quadrático interpolador da função dada na tabela.

Exercício 5.45 Determine uma aproximação de $\cos\frac{\pi}{8}$ usando o polinómio interpolador segmentado de grau 2 em 5 pontos no intervalo $[0,\pi]$ e indique um estimativa para o erro cometido. Compare esta estimativa com a obtida no Exercício 5.26.

Exercício 5.46 Determine polinómios interpoladores segmentados de grau 1 e 2 para a função $f(x)=x^3$ no intervalo [-1, 1].

Exercício 5.47 Determine o polinómio interpolador de Hermite de grau mínimo para a função $f(x)=\cos x, x\in[0,\frac{\pi}{2}]$ e calcule um valor aproximado para $\cos\frac{\pi}{8}$ e para $\sin\frac{\pi}{8}$.

Exercício 5.48 Determine o polinómio de grau mínimo que seja concordante[1] com a recta $y=-2+\frac{1}{2}(8-x)$, no ponto $(8,-2)$, e com a circunferência $(x-1)^2+(y+2)^2=1$, no ponto $(1,-1)$.

Exercício 5.49 Da função $f(x)=\sinh(x)=(e^x-e^{-x})/2$ conhecem-se os seguintes valores tabelados:

x_i	0	1
$f(x_i)$	0	$\frac{e-1/e}{2}$
$f'(x_i)$	1	$\frac{e+1/e}{2}$

.

1. A partir dos valores dados, calcule o valor aproximado de $f(0{,}5)$, usando interpolação polinomial cúbica adequada. ($e=2{,}71828182845905\ldots$)

2. Obtenha um majorante para o erro absoluto da aproximação obtida na alínea anterior (sem calculadora, obviamente).

Exercício 5.50 A tabela a seguir mostra a distância percorrida por um veículo, em função do tempo

t (s)	0	10	20	30	40	50	60
y (m)	0	50	125	205	280	350	410

.

1. Use um polinómio interpolador de grau 2 e os dados da tabela para estimar a distância percorrida pelo veículo ao fim de 45 segundos.

2. Seja $v(t)$ a velocidade do veículo no instante t, $v=\dfrac{dy}{dt}$. Sabendo que $v(30)=8,2$ e $v(60)=5,2$, calcule o polinómio de Hermite de grau mínimo e determine uma aproximação para $v(45)$.

Exercício 5.51 Considere um polinómio de grau 2, P, do qual se conhece a informação reunida na seguinte tabela:

x_i	-2	-1	-0,5	0
$P(x_i)$	1,8	2	α	3

.

1. Obtenha a expressão de P e calcule o valor de α.

2. Suponha que P interpola uma função $f:[-2,0]\longrightarrow\mathbb{R}$ nos pontos dados na tabela. Sabendo que $f'(-2)=1$ e $f'(0)=-1$, use um polinómio de Hermite para determinar uma aproximação para $f'(-0,5)$.

[1] Duas curvas dizem-se concordantes se tiverem a mesma tangente no ponto de união.

Exercício 5.52 Determine um polinómio que passa pelos pontos (-1,2) e (1,1) de modo a que o declive das rectas tangentes ao polinómio nesses pontos seja 0.

Exercício 5.53 Considere a função $f(x) = \sin x$, definida em $[0, \frac{\pi}{2}]$.

1. Determine quantos pontos deve considerar no intervalo dado para que o erro absoluto da aproximação de $f(x)$ por um polinómio interpolador de Hermite seja inferior a 0,1, para todo o $x \in [0, \frac{\pi}{2}]$.

2. Sabendo que $\sin(\frac{\pi}{3}) = \frac{\sqrt{3}}{2}$, determine uma aproximação para $\sqrt{3}$ utilizando o polinómio obtido na alínea anterior.

Exercício 5.54 (Matlab) De uma função f conhecem-se os valores dados na seguinte tabela:

x	-3	-2	-1	0	1	2	3
f	-1	-1	-1	0	1	1	1

.

Determine:

1. o polinómio interpolador de Lagrange nos referidos pontos;

2. o polinómio interpolador cúbico segmentado de Hermite (use o comando `pchip`);

3. o spline cúbico de interpolação (use o comando `spline`).

Compare os gráficos das funções obtidas nas alíneas anteriores.

Exercício 5.55 (Matlab) Considere a função $f(x) = \sin(2\pi x)$ em 21 nós equidistantes, x_i, $i = 1, 2, ..., 21$, no intervalo $[-1, 1]$. Determine:

1. o polinómio interpolador de Lagrange nos referidos pontos;

2. o polinómio interpolador cúbico segmentado de Hermite (use o comando `pchip`);

3. o spline cúbico de interpolação (use o comando `spline`).

Compare o gráfico das funções obtidas com o de f e repita os cálculos anteriores usando o seguinte conjunto de dados perturbados: $f(x_i) = \sin(2\pi x_i) + (-1)^{i+1}10^{-4}$, $i = 1, 2, ..., 21$.

Exercício 5.56 (Matlab) Considera-se um teste mecânico para estabelecer a relação entre tensões e deformações relativas a uma amostra de tecido biológico. Partindo dos valores da tabela

Teste	1	2	3	4	5	6	7	8
Tensão	0,00	0,06	0,14	0,25	0,31	0,47	0,60	0,70
Deformação	0,00	0,08	0,14	0,20	0,23	0,25	0,28	0,29

:

1. determine o polinómio interpolador;

2. obtenha o spline cúbico natural;

3. estime o valor da deformação correspondente a uma tensão igual a 0,9.

Exercício 5.57 (Matlab) Pretende-se aproximar a trajectória plana de um robot (idealizado como um ponto material) durante um ciclo de trabalho numa indústria. O robot deverá satisfazer algumas restrições: estar parado no ponto do plano $(0,0)$ no instante inicial $(t=0)$, deslocar-se até ao ponto $(1,2)$ em $t=1$, atingir o ponto $(4,4)$ em $t=2$, parar e iniciar de novo o movimento para atingir o ponto $(3,1)$ em $t=3$, voltar à sua posição inicial em $t=5$, parar e recomeçar um novo ciclo de trabalho. Para encontrar a trajectória do robot, proceda do seguinte modo: divida o intervalo de tempo $[0,5]$ em dois subintervalos $[0,2]$ e $[2,5]$. Em cada um dos subintervalos obtenha um spline que interpole os dados e tenha derivada nula nos extremos. Trace o respectivo gráfico.

5.7.2 Exercícios de aplicação à engenharia

Exercício 5.58 Conhecem-se as coordenadas de cinco pontos de uma curva plana que representa uma região de uma peça em corte. Determine o polinómio de Lagrange de grau 4 que interpola a referida curva sabendo que os pontos de coordenadas conhecidas são: $P_1=(1,2)$, $P_2=(2,1)$, $P_3=(3,1)$, $P_4=(4,2{,}5)$ e $P_5=(5,4)$. Determine ainda valores aproximados para as ordenadas dos pontos cujas abcissas são 0, $2{,}5$ e 6.

Exercício 5.59 Na seguinte tabela são dados diferentes valores para o peso específico p da água a diferentes temperaturas T (em graus centígados):

T	0	1	2	3
p	$0{,}999871$	$0{,}999928$	$0{,}999969$	$0{,}999991$

.

Usando interpolação linear, quadrática e cúbica, determine uma aproximação para p quando $T=4\ ^o$C. Compare os resultados obtidos sabendo que o valor exacto é $1{,}000000$.

Exercício 5.60 Durante a sedimentação da reacção de saponificação entre quantidades equimolares de hidróxido de sódio e acetato de etilo, a concentração c (gr mole/litro) de cada reagente varia com o tempo t (min) de acordo com a equação

$$\frac{1}{c}=\frac{1}{c_0}+kt,$$

onde c_0 é a concentração inicial e k (litro/gr mole min) é a constante de reacção. Foram obtidos os seguintes resultados em laboratório à temperatura de $77\ ^o$F:

$1/c$	$24{,}7$	$32{,}4$	$38{,}4$	$45{,}0$	$52{,}3$	$65{,}6$	$87{,}6$	102	154	192
t	1	2	3	4	5	7	10	12	20	25

.

1. Obtenha uma estimativa para a concentração inicial.
2. Obtenha uma estimativa para a concentração ao fim de 15 minutos e compare-a com a solução obtida em laboratório (ao fim de 15 minutos obteve-se $1/c=135$).

Exercício 5.61 O censo da população dos Estados Unidos, entre 1930 e 1980, produziu os seguintes resultados:

Ano	1930	1940	1950	1960	1970	1980
População ($\times 10^3$)	123203	131669	150697	179323	203212	226505

.

Use um polinómio interpolador apropriado para estimar a população nos anos de 1920, 1965, e 2000. Sabendo que a população no ano de 1920 era de 105711×10^3, o que pode inferir quanto à precisão das aproximações obtidas para os anos de 1965 e 2000?

Exercício 5.62 Determine uma aproximação para o instante da passagem do perigeu da Lua em Março de 1999, a partir dos valores tabelados para as zero horas de cada dia

dia	19	20	21
distância	57,071	56,955	57,059

.

Indique também a distância (em raios médios da Terra) da Terra à Lua nesse instante.

Exercício 5.63 Usando interpolação cúbica livre, determine uma aproximação para a declinação aparente de Vénus para o dia 9 de Maio de 1999, às 18h30m45s, a partir das Efemérides Astronómicas (onde está tabelada para cada dia, às zero horas)

dia	7	8	9	10
δ_i	$+5^o51'47''$,55	$+6^o22'25''$,20	$+6^o52'54''$,57	$+6^o23'14''$,96

.

A partir da função obtida, determine uma aproximação para o instante em que a declinação aparente de Vénus no dia 9 de Maio de 1999 foi máxima.

Exercício 5.64 A estrela S da Ursa Maior apresenta uma variação para a sua magnitude aparente m, em função do ângulo de fase θ (em graus), de acordo com os dados da seguinte tabela:

θ	-60	-20	20
m	9,40	11,39	10,84

.

Usando um *spline* cúbico natural, determine uma aproximação para o ângulo de fase pertencente ao intervalo $[-20, 20]$ em que a magnitude aparente da estrela é máxima.

Exercício 5.65 Deslocando-se um receptor de GPS num veículo ao longo do eixo de uma estrada, em Angola, obtiveram-se as coordenadas locais:

latitude ϕ	$26'56''$,1	$26'50''$,4	$27'02''$,7	$26'58''$,3
longitude λ	$5'36''$	$5'56''$	$6'16''$	$6'36''$

.

Aproximando o eixo da estrada por um *spline* natural determine:

1. a latitude da estrada quando a longitude é $\lambda = 6'$;
2. as coordenadas da estrada no ponto mais perto do equador, supondo que isso acontece entre $6'16''$ e $6'36''$ de longitude.

Exercício 5.66 Um carro percorre uma rua, em linha recta, tendo sido efectuados os seguintes registos:

tempo (t) em segundos	0	5	10
distância (d) em metros	0	90	150
velocidade (v) em km/hora	40		40

.

Usando o *spline* cúbico completo interpolador da função distância nos pontos dados, indique, justificando, uma aproximação para:

1. o primeiro instante em que o carro excedeu o limite de velocidade permitido dentro das localidades;
2. o instante em que o carro atingiu a velocidade máxima nos primeiros 5 segundos.

Exercício 5.67 Uma das formas mais utilizadas na construção de curvas consiste em partir das respectivas equações paramétricas e proceder a uma interpolação apropriada. Considere o caso das curvas planas dadas pelas equações paramétricas

$$\left\{ \begin{array}{rcl} x & = & p(t) \\ y & = & q(t) \end{array} \right. , \qquad t \in [0,1],$$

em que p e q são polinómios.

1. Determine a forma destes polinómios de modo a que a curva passe pelos pontos $P_0 = (x_0, y_0)$ e $P_1 = (x_1, y_1)$ com tangentes $T_0 = (p'(0), q'(0))$ e $T_1 = (p'(0), q'(0))$, respectivamente (curva de Ferguson-Coons).

2. A especificação das tangentes através de dois pontos auxiliares (pontos de guia) revela-se mais útil na prática. Assim, sejam P_2 e P_3 dois pontos auxiliares tais que $T_0 = \lambda(P_2 - P_0) = \lambda(x_2 - x_0, y_2 - y_0)$ e $T_1 = \lambda(P_3 - P_1) = \lambda(x_3 - x_1, y_3 - y_1)$, em que λ é um factor de normalização à escolha. Mostre que a curva de Ferguson-Coons se pode escrever na forma

$$\left\{ \begin{array}{rcl} x(t) & = & \phi_0(t)x_0 + \phi_1(t)x_1 + \phi_2(t)x_2 + \phi_3(t)x_3 \\ y(t) & = & \phi_0(t)y_0 + \phi_1(t)y_1 + \phi_2(t)y_2 + \phi_3(t)y_3 \end{array} \right. , \qquad t \in [0,1],$$

   com

$$\begin{array}{rcl} \phi_0(t) & = & 2t^3 - \lambda t^3 - 3t^2 + 2\lambda t^2 - \lambda t + 1, \\ \phi_1(t) & = & -2t^3 + \lambda t^3 - \lambda t^2 + 3t^2, \\ \phi_2(t) & = & \lambda t^3 - 2\lambda t^2 + \lambda t, \\ \phi_3(t) & = & -\lambda t^3 + \lambda t^2. \end{array}$$

3. Mostre que a curva está contida no invólucro convexo definido pelos pontos P_0, P_1, P_2 e P_3, isto é, que $\phi_i(t) \geq 0, i = 0, 1, 2, 3$, e que $\sum_{i=0}^{3} \phi_i(x) = 1$, se e só se $0 \leq \lambda \leq 3$.

   Nota: Quando $\lambda = 3$ a curva de Ferguson-Coons também se chama curva de Pierre Étienne Bézier (1910-1999).

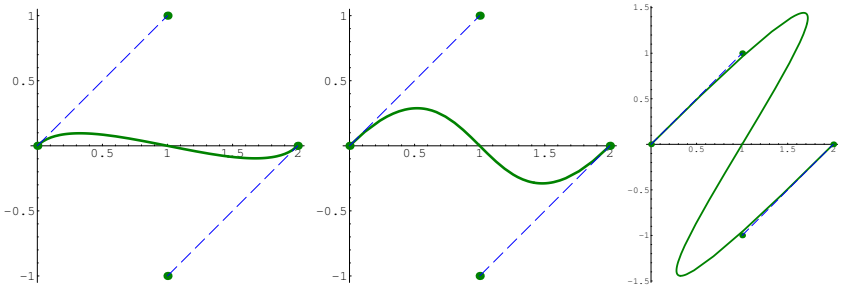

Figura 5.1: Gráficos das curvas de Ferguson-Coons com $\lambda = 1, 3$ e 15.

Capítulo 6

Derivação e integração numérica

6.1 Derivação numérica

Acontece frequentemente sermos confrontados com a necessidade de determinar valores da derivada de uma função num conjunto de pontos conhecendo o valor da função apenas nesses pontos. Na impossibilidade de obter esses valores de forma exacta, vamos considerar a sua aproximação através do valor da derivada do polinómio interpolador da função nos referidos pontos.

Para o estudo que iremos efetuar, consideremos uma função $f \in C^m([a,b])$, com m suficientemente grande por forma a que as deduções das fórmulas possam ser efectuadas, conhecida num conjunto de pontos da partição uniforme

$$a = x_0 < x_1 < \cdots < x_n = b, \tag{6.1}$$

com $x_k - x_{k-1} = h$, $k = 1, \ldots, n$.

6.1.1 Aproximação da primeira derivada

Queremos aproximar a derivada de $f \in C^m([a,b])$ num dos pontos x_k, $k \in \{0, 1, \ldots, n\}$, da partição (6.1).

Fórmulas com dois pontos

Considerando a fórmula de Taylor para f em torno do ponto x_k, e assumindo que $f \in C^2([a,b])$, temos

$$f(x_{k+1}) = f(x_k) + f'(x_k)h + \frac{h^2}{2}f''(\xi_1), \qquad \xi_1 \in]x_k, x_{k+1}[$$

e

$$f(x_{k-1}) = f(x_k) - f'(x_k)h + \frac{h^2}{2}f''(\xi_2), \qquad \xi_2 \in]x_{k-1}, x_k[.$$

Assim sendo, podemos escrever

$$f'(x_k) = \frac{f(x_{k+1}) - f(x_k)}{h} - \frac{h}{2}f''(\xi_1), \qquad \xi_1 \in]x_k, x_{k+1}[$$

e

$$f'(x_k) = \frac{f(x_k) - f(x_{k-1})}{h} + \frac{h}{2}f''(\xi_2), \qquad \xi_2 \in]x_{k-1}, x_k[.$$

Obtemos assim duas fórmulas de diferenças finitas de primeira ordem em h para aproximar a primeira derivada de uma função num ponto. A

$$(\delta_+ f)(x_k) = \frac{f(x_{k+1}) - f(x_k)}{h}$$

é usual chamar fórmula de diferenças finitas progressivas e a

$$(\delta_- f)(x_k) = \frac{f(x_k) - f(x_{k-1})}{h}$$

costuma chamar-se fórmula de diferenças finitas regressivas.

Fórmulas com três pontos

Para obter fórmulas mais precisas para aproximar a primeira derivada de uma função num ponto, vamos considerar fórmulas com mais pontos. No próximo exercício apresentam-se as fórmulas de diferenças finitas progressivas, centradas e regressivas com três pontos.

Exercício 6.1 Prove que:

1. $f'(x_k) = \dfrac{1}{2h}\left[-3f(x_k) + 4f(x_{k+1}) - f(x_{k+2})\right] + \dfrac{h^2}{3}f'''(\xi), \qquad \xi \in]x_k, x_{k+2}[;$
2. $f'(x_k) = \dfrac{1}{2h}\left[f(x_{k+1}) - f(x_{k-1})\right] - \dfrac{h^2}{6}f'''(\xi), \qquad \xi \in]x_{k-1}, x_{k+1}[;$
3. $f'(x_k) = \dfrac{1}{2h}\left[f(x_{k-2}) - 4f(x_{k-1}) + 3f(x_k)\right] + \dfrac{h^2}{3}f'''(\xi), \qquad \xi \in]x_{k-2}, x_k[.$

Resolução: Vamos só deduzir a segunda fórmula. Desenvolvendo f em série de Taylor em torno do ponto x_k, e assumindo que $f \in C^3([a,b])$, temos

$$f(x_{k+1}) = f(x_k) + f'(x_k)h + \frac{h^2}{2}f''(x_k) + \frac{h^3}{6}f'''(\xi_1), \qquad \xi_1 \in]x_k, x_{k+1}[;$$

$$f(x_{k-1}) = f(x_k) - f'(x_k)h + \frac{h^2}{2}f''(x_k) - \frac{h^3}{6}f'''(\xi_2), \qquad \xi_2 \in]x_{k-1}, x_k[.$$

Subtraindo membro a membro, e colocando $f'(x_k)$ em evidência, obtemos

$$f'(x_k) = \frac{1}{2h}[f(x_{k+1}) - f(x_{k-1})] - h^2\frac{f'''(\xi_1) + f'''(\xi_2)}{12}, \qquad \xi_1 \in]x_k, x_{k+1}[, \xi_2 \in]x_{k-1}, x_k[.$$

Como f''' é contínua em $[a,b]$, existe um $\xi \in]x_{k-1}, x_{k+1}[$ tal que

$$f'''(\xi) = \frac{1}{2}\left(f'''(\xi_1) + f'''(\xi_2)\right). \tag{6.2}$$

De facto, atendendo a que f''' é contínua e o intervalo $[a,b]$ é fechado, temos que

$$2\min_{x\in[a,b]} f'''(x) \leq f'''(\xi_1) + f'''(\xi_2) \leq 2\max_{x\in[a,b]} f'''(x).$$

Pelo Teorema de Bolzano conclui-se que existe $\xi \in]x_{k-1}, x_{k+1}[$ tal que (6.2) se verifica. Provámos, assim, que

$$f'(x_k) = \frac{1}{2h}[f(x_{k+1}) - f(x_{k-1})] - h^2\frac{f'''(\xi)}{6}, \qquad \xi \in]x_{k-1}, x_{k+1}[.$$

A

$$(\delta f)(x_k) = \frac{1}{2h}[f(x_{k+1}) - f(x_{k-1})]$$

chamamos fórmula de diferenças finitas centradas de segunda ordem em h.

O próximo exercício generaliza o raciocínio efectuado no exercício anterior e vai ser usado, frequentemente, para deduzir as fórmulas do erro na derivação e na integração numérica.

Exercício 6.2 Mostre que, se f for uma função contínua em $[a, b]$ e $\xi_k \in [a, b]$, para $k = 1, \ldots, M$, então existe um $\xi \in [a, b]$ tal que

$$f(\xi) = \frac{1}{M} \sum_{k=1}^{M} f(\xi_k).$$

Exercício 6.3 Considere os seguintes valores da função $f(x) = xe^x$:

x_i	1,8	1,9	2,0	2,1	2,2
$f(x_i)$	10,889365	12,703199	14,778112	17,148957	19,855030

.

Aproxime o valor de $f'(2{,}0) = 22{,}167168$ usando as fórmulas de diferenças finitas dadas no exercício anterior e compare os erros cometidos.

Resolução: Vamos considerar as três fórmulas separadamente.

- Fórmula progressiva de segunda ordem com $h = 0.1$.

  $$f'(2{,}0) \approx \frac{1}{0{,}2}[-3f(2{,}0) + 4f(2{,}1) - f(2{,}2)] = 22{,}032310.$$

  O erro cometido é aproximadamente $1{,}35 \times 10^{-1}$.

- Fórmula regressiva de segunda ordem com $h = 0{,}1$.

  $$f'(2{,}0) \approx \frac{1}{0{,}2}[f(1{,}8) - 4f(1{,}9) + 3f(2{,}0)] = 22{,}054525.$$

  O erro cometido é aproximadamente $1{,}13 \times 10^{-1}$.

- Fórmula centrada de segunda ordem com $h = 0{,}1$.

  $$f'(2{,}0) \approx \frac{1}{0{,}2}[f(2{,}1) - f(1{,}9)] = 22{,}228790.$$

  O erro cometido é aproximadamente $-6{,}16 \times 10^{-2}$.

Note-se que o erro cometido quando se usa a fórmula de diferenças centradas é aproximadamente metade do erro cometido com as outras fórmulas, o que confirma o resultado do exercício anterior.

6.1.2 Aproximação da segunda derivada

Queremos aproximar a segunda derivada de f num dos pontos x_k, $k \in \{0, 1, \ldots, n\}$, da partição (6.1). Desenvolvendo f em série de Taylor em torno do ponto x_k, e assumindo que $f \in C^4([a, b])$, temos

$$f(x_{k+1}) = f(x_k) + f'(x_k)h + \frac{h^2}{2}f''(x_k) + \frac{h^3}{6}f'''(x_k) + \frac{h^4}{24}f^{(4)}(\xi_1), \qquad \xi_1 \in]x_k, x_{k+1}[$$

e

$$f(x_{k-1}) = f(x_k) - f'(x_k)h + \frac{h^2}{2}f''(x_k) - \frac{h^3}{6}f'''(x_k) + \frac{h^4}{24}f^{(4)}(\xi_2), \qquad \xi_2 \in]x_{k-1}, x_k[.$$

Se adicionarmos estas duas expressões obtemos

$$f''(x_k) = \frac{1}{h^2}[f(x_{k-1}) - 2f(x_k) + f(x_{k+1})] - \frac{h^2}{24}\left(f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right).$$

Uma vez que $f^{(4)}$ é contínua em $[x_{k-1}, x_{k+1}]$, o Exercício 6.2 permite concluir que existe um $\xi \in]x_{k-1}, x_{k+1}[$ tal que

$$f^{(4)}(\xi) = \frac{1}{2}\left(f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right).$$

Assim,

$$f''(x_k) = \frac{1}{h^2}[f(x_{k-1}) - 2f(x_k) + f(x_{k+1})] - \frac{h^2}{12}f^{(4)}(\xi). \tag{6.3}$$

A fórmula

$$(\delta_2 f)(x_k) = \frac{1}{h^2}[f(x_{k-1}) - 2f(x_k) + f(x_{k+1})] \tag{6.4}$$

é conhecida como fórmula de diferenças centradas de segunda ordem para aproximar a segunda derivada. Por um raciocínio semelhante poderiam ser obtidas outras fórmulas de diferenças finitas para aproximar a segunda derivada, não só centradas como também progressivas e regressivas.

Exercício 6.4 Prove que

$$f''(x_k) = \frac{1}{12h^2}[-f(x_{k-2}) + 16f(x_{k-1}) - 30f(x_k) + 16f(x_{k+1}) - f(x_{k+2})] + \frac{h^4}{90}f^{(6)}(\xi),$$

com $\xi \in]x_{k-2}, x_{k+2}[$.

Exercício 6.5 Considere, de novo, os valores da função $f(x) = xe^x$ dados na tabela do Exercício 6.3. Aproxime o valor de $f''(2{,}0) = 29{,}556224$ usando a fórmula de diferenças finitas centradas de segunda ordem.

Resolução: Temos que

$$f''(2{,}0) \approx \frac{1}{0{,}01}[f(1{,}9) - 2f(2{,}0) + f(2{,}1)] = 29{,}593200.$$

O erro cometido é aproximadamente $-3{,}7 \times 10^{-2}$.

Exercício 6.6 Mostre, a partir do polinómio interpolador de Lagrange da função f nos pontos x_0, x_1 e x_2, tais que $x_1 - x_0 = h$ e $x_2 - x_1 = \alpha h$, que

$$f''(x) \approx \frac{2}{h^2}\left[\frac{f(x_0)}{1+\alpha} - \frac{f(x_1)}{\alpha} + \frac{f(x_2)}{\alpha(1+\alpha)}\right].$$

Verifique que quando $\alpha = 1$ se recupera a fórmula das diferenças centradas.

6.2 Integração numérica

Nesta secção vamos obter e analisar as chamadas fórmulas de quadratura numérica que permitem determinar, de forma aproximada, o integral definido

$$I(f) = \int_a^b f(x)dx$$

de uma função real de variável real f num dado intervalo real $[a, b]$.

Seja f uma função conhecida em $M+1$ pontos $a = x_0 < x_1 < \cdots < x_{M-1} < x_M = b$, com $x_k = x_0 + kH$ e $H = (b-a)/M$. Assim sendo, temos que

$$I(f) = \int_a^b f(x)dx = \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} f(x)dx.$$

As chamadas fórmulas que iremos considerar permitem obter aproximações a $I(f)$ aproximando, em cada intervalo $I_k = [x_{k-1}, x_k]$, $k = 1, \ldots, M$, a função f por um seu polinómio interpolador. Se P_n^k for o polinómio de grau menor ou igual a n interpolador de f em $n+1$ pontos de I_k, temos que $f(x) = P_n^k(x) + e_n^k(x)$, com $x \in I_k$, onde $e_n^k(x)$ é o erro cometido na interpolação. Assim,

$$I(f) = I_{NC}(f; n) + E_{NC}(f; n),$$

com

$$I_{NC}(f; n) = \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} P_n^k(x)dx \quad \text{e} \quad E_{NC}(f; n) = \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} e_n^k(x)dx. \tag{6.5}$$

As fórmulas (6.5) são conhecidas como fórmulas de Newton-Cotes (compostas), em homenagem a Newton e Roger Cotes (1682-1716), e dependem, obviamente, do grau do polinómio escolhido. As fórmulas de Newton-Cotes (simples) são aquelas que se obtêm quando se considera $M = 1$ ou, de forma equivalente, $H = b - a$.

Uma vez que a n-ésima fórmula de Newton-Cotes é obtida à custa da aproximação da função integranda por um polinómio de grau n, será de esperar que esta seja exacta para polinómios de grau menor ou igual a n. Este facto conduz-nos ao conceito de grau de exactidão de uma fórmula de quadratura numérica.

Definição 6.1 (Grau de exactidão) *Uma fórmula de quadratura numérica (6.5) diz-se com* grau de exactidão n *se é exacta para polinómios de grau menor ou igual a* n.

Outra forma de analisar o erro cometido ao aproximar um integral por uma fórmula de quadratura numérica é através da sua ordem de convergência.

Definição 6.2 (Ordem de convergência) *Uma fórmula de quadratura numérica (6.5) diz-se com* ordem de convergência p *se* $|E_{NC}(f; n)| \leq CH^p$, *com* C *uma constante independente de* H *e* M.

6.2.1 Fórmula do ponto médio

Consideremos, em I_k, $k = 1, \ldots, M$, a função f aproximada pelo seu polinómio interpolador de grau 0 no ponto médio do intervalo $\overline{x}_k = (x_{k-1} + x_k)/2$, isto é,

$$f(x) \approx f(\overline{x}_k), \qquad x \in I_k.$$

Para isso, a função f também terá que ser conhecida nos pontos $\overline{x}_k$, $k = 1, ..., M$. Então,

$$I(f) = \int_a^b f(x)dx = \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} f(x)dx \approx \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} f(\overline{x}_k)dx = H \sum_{k=1}^{M} f(\overline{x}_k).$$

Obtivemos, assim, a chamada fórmula do ponto médio (composta)

$$I^c_{PM}(f) = H \sum_{k=1}^{M} f(\overline{x}_k).$$

A fórmula do ponto médio (simples) é a que se obtém quando se considera $M = 1$, isto é,

$$I_{PM}(f) = (b-a) f\left(\frac{a+b}{2}\right).$$

Vamos agora analisar o erro que se comete na aproximação $I(f) \approx I_{PM}(f)$. Para isso, temos que considerar o seguinte teorema.

Teorema 6.1 (Valor Médio para integrais) *Seja f uma função contínua em $[a, b]$ e g uma função integrável que não muda de sinal em $[a, b]$. Então existe pelo menos um $\xi \in]a, b[$ tal que*

$$f(\xi) \int_a^b g(x)dx = \int_a^b f(x)g(x)dx.$$

Usando o desenvolvimento em série de Taylor, e assumindo que $f \in C^2([a, b])$, temos que, considerando $\overline{x} = (a + b)/2$,

$$\begin{aligned} I(f) - I_{PM}(f) &= \int_a^b (f(x) - f(\overline{x}))\, dx \\ &= \int_a^b \left(f'(\overline{x})(x - \overline{x}) + \frac{f''(\xi_x)}{2}(x - \overline{x})^2 \right) dx, \quad \xi_x \in I\{x, \overline{x}\}. \end{aligned}$$

Atendendo a que (prove)

$$\int_a^b (x - \overline{x})dx = 0$$

e ao facto de $(x - \overline{x})^2$ não mudar de sinal em $[a, b]$, usando o Teorema 6.1, temos que

$$I(f) - I_{PM}(f) = \frac{f''(\xi)}{2} \int_a^b (x - \overline{x})^2 dx, \qquad \xi \in]a, b[.$$

Então (prove)

$$E_{PM}(f) = I(f) - I_{PM}(f) = \frac{f''(\xi)}{24}(b - a)^3, \qquad \xi \in]a, b[.$$

Pelas definições dadas anteriormente, temos que a fórmula do ponto médio simples tem ordem de convergência 3 e grau de exactidão 1. No entanto, como veremos no próximo exercício, a fórmula do ponto médio perde uma ordem de convergência quando usada na sua forma composta.

Exercício 6.7 Mostre que

$$E^c_{PM}(f) = I(f) - I^c_{PM}(f) = \frac{H^2}{24}(b - a)f''(\xi), \qquad \xi \in]a, b[.$$

Resolução: Temos que

$$E_{PM}^c(f) = \sum_{k=1}^{M} \frac{f''(\xi_k)}{24}(x_{k-1} - x_k)^3 = \frac{H^3}{24}\sum_{k=1}^{M} f''(\xi_k), \qquad \xi_k \in]x_{k-1}, x_k[.$$

Atendendo ao Exercício 6.2 provamos o pretendido.

Face ao exercício anterior, podemos afirmar que a fórmula do ponto médio tem ordem de convergência 2 e grau de exactidão 1.

Na prática a fórmula do erro aparece, normalmente, em valor absoluto. É usual considerar a expressão

$$|E_{PM}^c(f)| \leq \frac{H^2}{24}(b-a)M_2, \quad \text{com} \quad M_2 = \max_{x\in[a,b]} |f''(x)|.$$

6.2.2 Fórmula do trapézio

Vamos considerar o caso em que pretendemos aproximar, em cada intervalo I_k, $k = 1, \ldots, M$, uma função $f \in C^2([a,b])$ por um polinómio do primeiro grau que passa pelos pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$. Como é sabido, usando interpolação de Lagrange,

$$f(x) = f(x_{k-1})\frac{x - x_k}{x_{k-1} - x_k} + f(x_k)\frac{x - x_{k-1}}{x_k - x_{k-1}} + \frac{f''(\xi_k)}{2}(x - x_{k-1})(x - x_k),$$

com $\xi_k \in]x_{k-1}, x_k[$ um valor que depende de x. Assim,

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{k=1}^{M}\int_{x_{k-1}}^{x_k} f(x)dx \approx \sum_{k=1}^{M}\int_{x_{k-1}}^{x_k}\left(f(x_{k-1})\frac{x - x_k}{x_{k-1} - x_k} + f(x_k)\frac{x - x_{k-1}}{x_k - x_{k-1}})\right)dx \\ &= \frac{H}{2}\sum_{k=1}^{M}(f(x_{k-1}) + f(x_k)) = \frac{H}{2}[f(a) + 2f(x_1) + \cdots + 2f(x_{M-1}) + f(b)].\end{aligned}$$

Obtivemos, assim, a chamada fórmula do trapézio (composta)

$$I_T^c(f) = \frac{H}{2}[f(a) + 2f(x_1) + \cdots + 2f(x_{M-1}) + f(b)].$$

A fórmula do trapézio (simples) é a que se obtém quando se considera $M = 1$, isto é,

$$I_T(f) = \frac{b-a}{2}[f(a) + f(b)]. \tag{6.6}$$

Vamos agora analisar o erro que se comete na aproximação $I(f) \approx I_T(f)$. Temos que, se $f \in C^2([a,b])$,

$$I(f) - I_T(f) = \int_a^b \frac{f''(\xi_x)}{2}(x-a)(x-b), \qquad \xi_x \in]a,b[.$$

Como $(x-a)(x-b)$ não muda de sinal em $]a,b[$, pelo Teorema 6.1, temos que (prove)

$$E_T(f) = I(f) - I_T(f) = -\frac{(b-a)^3}{12}f''(\xi), \qquad \xi \in]a,b[.$$

Tal como para a fórmula do ponto médio simples, temos que a fórmula do trapézio simples tem ordem de convergência 3 e grau de exactidão 1. No entanto, como veremos no próximo exercício, a fórmula do trapézio, tal como a do ponto médio, perde uma ordem de convergência quando usada na sua forma composta.

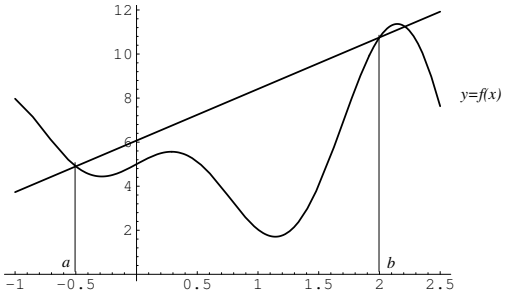

Figura 6.1: Fórmula do trapézio.

Exercício 6.8 Mostre que

$$E_T^c(f) = I(f) - I_T^c(f) = -\frac{H^2}{12}(b-a)f''(\xi), \qquad \xi \in]a,b[.$$

Resolução: Temos que

$$E_T^c(f) = -\sum_{k=1}^{M} \frac{f''(\xi_k)}{12}(x_{k-1}-x_k)^3 = -\frac{H^3}{12}\sum_{k=1}^{M} f''(\xi_k), \qquad \xi_k \in]x_{k-1}, x_k[.$$

Atendendo ao Exercício 6.2 provamos o pretendido.

Face ao exercício anterior, podemos afirmar que a fórmula do trapézio, tal como a do ponto médio, tem ordem de convergência 2 e grau de exactidão 1.

Na prática a fórmula do erro aparece, normalmente, em valor absoluto. É usual considerar a expressão

$$|E_T^c(f)| \leq \frac{H^2}{12}(b-a)M_2, \quad \text{com} \quad M_2 = \max_{x\in[a,b]} |f''(x)|.$$

O valor do integral de uma determinada função f num intervalo $[a,b]$ pela fórmula do trapézio pode ser dado de acordo com o seguinte algoritmo.

Algoritmo 6.1 Fórmula do trapézio

Dados: a, b e M

$H := (b-a)/M$

$x := a$

$s := 0$

Para k de 1 até $M-1$ fazer

$\quad x := x + H$

$\quad s := s + f(x)$

$I_T := (H/2)(f(a) + 2s + f(b))$

Resultado: $I \approx I_T$

Exercício 6.9 Seja $I = \int_{-2}^{-1} xe^{2x}dx$. Calcule, usando a fórmula do trapézio, o valor aproximado de I com três casas decimais correctas.

Resolução: Seja $f(x) = xe^{2x}$. Temos que, para $x \in [-2,-1]$, o erro para a fórmula do trapézio é dado por

$$|E_T^c(x)| \leq \frac{1}{12}H^2M_2 = \frac{1}{12M^2}M_2,$$

sendo

$$M_2 = \max_{x\in[-2,-1]}|f''(x)| = \max_{x\in[-2,-1]}(-4e^{2x}(x+1)).$$

Se tomarmos $g(x) = -4e^{2x}(x+1)$ temos que $g'(x) = 0 \Rightarrow x = -1{,}5$. Logo

$$M_2 = \max\{g(-2), g(-1{,}5), g(-1)\} = 2e^{-3}.$$

Vamos então determinar qual o menor valor de M que satisfaz

$$\frac{e^{-3}}{6M^2} \leq 0{,}5 \times 10^{-3}.$$

Efectuando os cálculos, concluímos imediatamente que $M \geq 4{,}074$ o que implica $M = 5$. Necessitamos de 6 pontos igualmente distanciados no intervalo $[-2,-1]$ para obter uma aproximação ao valor de I com três casas decimais correctas. Assim,

$$I \approx 0{,}1[f(-2)+2f(-1{,}8)+2f(-1{,}6)+2f(-1{,}4)+2f(-1{,}2)+f(-1)] = -0{,}0788762.$$

Como só podemos garantir três casas decimais correctas temos que $I \approx -0{,}079$.

6.2.3 Fórmula de Simpson

Consideremos agora o caso em que pretendemos aproximar, em cada intervalo I_k, $k = 1,\ldots,M$, uma função $f \in C^3([a,b])$ por um polinómio do segundo grau que passa pelos pontos $(x_{k-1}, f(x_{k-1}))$, $(\overline{x}_k, f(\overline{x}_k))$ e $(x_k, f(x_k))$, com $\overline{x}_k = (x_{k-1}+x_k)/2$. Como foi visto no capítulo dedicado à interpolação de Lagrange, para $x \in I_k$,

$$\begin{aligned} f(x) &= f(x_{k-1})\frac{2(x-\overline{x}_k)(x-x_k)}{H^2} - f(\overline{x}_k)\frac{4(x-x_{k-1})(x-x_k)}{H^2} \\ &+ f(x_{k+1})\frac{2(x-x_{k-1})(x-\overline{x}_k)}{H^2} + \frac{f'''(\xi_k)}{6}(x-x_{k-1})(x-\overline{x}_k)(x-x_k), \end{aligned}$$

com $\xi_k \in]x_{k-1}, x_k[$ um valor que depende de x.

Exercício 6.10 Prove que o valor do integral

$$\int_{x_{k-1}}^{x_k}\left(f(x_{k-1})\frac{2(x-\overline{x}_k)(x-x_k)}{H^2} - f(\overline{x}_k)\frac{4(x-x_{k-1})(x-x_k)}{H^2} + f(x_{k+1})\frac{2(x-x_{k-1})(x-\overline{x}_k)}{H^2}\right)dx,$$

com $\overline{x}_k = (x_{k-1}+x_k)/2$, é dado por

$$\frac{H}{6}\left[f(x_{k-1}) + 4f(\overline{x}_k) + f(x_k)\right].$$

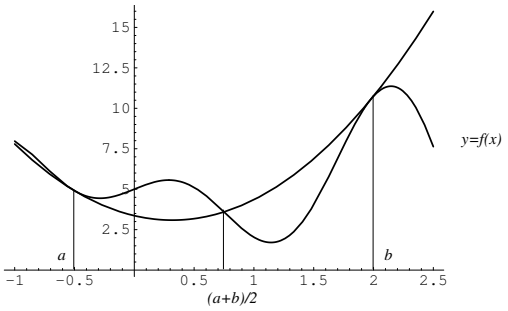

Figura 6.2: Fórmula de Simpson.

Pelo exercício anterior podemos obter a chamada fórmula de Simpson (composta)

$$I_S^c(f) = \frac{H}{6} \sum_{k=1}^{M} \left[f(x_{k-1}) + 4f(\overline{x}_k)) + f(x_k) \right],$$

estabelecida por Thomas Simpson (1710-1761), que também pode ser escrita na forma

$$I_S^c(f) = \frac{H}{6} \left[f(a) + 4 \sum_{k=1}^{M} f(\overline{x}_k) + 2 \sum_{k=1}^{M-1} f(x_k) + f(b) \right]. \tag{6.7}$$

A fórmula de Simpson (simples) é a que se obtém quando se considera $M = 1$, isto é,

$$I_S(f) = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right].$$

Ao contrário do que foi efectuado para as fórmulas do ponto médio e do trapézio, neste caso não podemos aplicar o Teorema 6.1 para determinar o erro cometido na aproximação $I(f) \approx I_S(f)$, uma vez que $(x-a)(x-(a+b)/2)(x-b)$ muda de sinal em $[a,b]$. É possível, no entanto, demonstrar que, se $f \in C^4([a,b])$, o erro associado à fórmula de Simpson (simples) é dado por

$$E_S(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \qquad \xi \in]a,b[.$$

Uma vez que a fórmula de Simpson foi obtida pela aproximação da função integranda por um polinómio de segundo grau, seria de esperar que tivesse grau de exactidão 2. No entanto, de forma surpreendente, a expressão obtida para o erro diz-nos que a fórmula de Simpson tem grau de exactidão três, isto é, esta fórmula é exacta sempre que a função a integrar é um polinómio de grau menor ou igual a 3.

A determinação do valor do erro que está associado à fórmula de Simpson (composta) pode ser feita de forma semelhante ao efectuado para a fórmula do trapézio (composta). De facto, pelo Exercício 6.2, como $f \in C^4([a,b])$, existe um $\xi \in]a,b[$ tal que

$$\sum_{k=1}^{M} f^{(4)}(\xi_k) = M f^{(4)}(\xi).$$

Assim sendo,

$$E_S^c(f) = -\frac{H^5}{2880} M f^{(4)}(\xi) = -\frac{H^4}{2880}(b-a) f^{(4)}(\xi), \qquad \xi \in]a,b[.$$

Na fórmula a prática é usual considerar a fórmula do erro em valor absoluto. No caso da fórmula de Simpson temos que

$$|E_S^c(f)| \leq \frac{H^4}{2880}(b-a)M_4, \quad \text{com} \quad M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|.$$

O exercício seguinte dá-nos uma forma alernativa de escrever a fórmula de Simpson.

Exercício 6.11 Seja f uma função conhecida apenas nos pontos $(x_k, f(x_k))$, $k = 0, 1, \ldots, M$, com M par, $x_k = a + kH$ e $H = (b-a)/M$, mostre que uma aproximação para o integral $\int_a^b f(x)dx$ pela a fórmula de Simpson (composta) é dado por

$$I_S^c(f) = \frac{H}{3}[f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{M-2}) + 4f(x_{M-1}) + f(b)], \quad (6.8)$$

com erro

$$E_S^c(f) = -\frac{H^4}{180}(b-a)f^{(4)}(\xi), \qquad \xi \in]a, b[,$$

o que implica

$$|E_S^c(f)| \leq \frac{H^4}{180}(b-a)M_4, \quad \text{com} \quad M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|. \quad (6.9)$$

A fórmula (6.8) é aquela que mais iremos usar na resolução de exercícios práticos. Nessa fórmula, os pontos x_k, com k ímpar, correspondem, em (6.7), aos pontos médios do intervalo.

O valor do integral de uma determinada função f num intervalo $[a, b]$ pela fórmula de Simpson pode ser dado de acordo com o seguinte algoritmo.

Algoritmo 6.2 Fórmula de Simpson

Dados: a, b e M (par)
$H := (b-a)/M$
$x := a$
$s := 0$
Para k de 1 até $M-1$ fazer
    $x := x + H$
    Se k par então $s := s + 2f(x)$ caso contrário $s := s + 4f(x)$
$I_S := (H/3)(f(a) + s + f(b))$
Resultado: $I \approx I_S$

Exercício 6.12 Seja $I = \int_0^1 e^x \cos x dx$. Calcule, usando a fórmula de Simpson, o valor aproximado de I com erro inferior a 10^{-3}.

Resolução: Seja $f(x) = e^x \cos x$. Temos que, para $x \in [0,1]$, o erro dado pela fórmula de Simpson (6.8) é (ver (6.9))

$$|E_S(x)| \leq \frac{1}{180} H^4 M_4 = \frac{1}{180M^4} M_4,$$

sendo

$$M_4 = \max_{x\in[0,1]} |f^{(4)}(x)| = \max_{x\in[0,1]} (4e^x \cos x).$$

Se tomarmos $g(x) = 4e^x \cos x$ temos que $g'(x) = 0 \Rightarrow x = \frac{\pi}{4}$. Logo

$$M_4 = \max\{g(0), g(\frac{\pi}{4}), g(1)\} = 2\sqrt{2}e^{\pi/4}.$$

Vamos então determinar qual o menor valor de M que satisfaz

$$\frac{\sqrt{2}e^{\pi/4}}{90M^4} < 10^{-3}.$$

Efectuando os cálculos, concluímos imediatamente que $M \geq 2{,}42$. Como M tem que ser par temos que $M = 4$. Então, necessitamos de 5 pontos igualmente distanciados no intervalo $[0,1]$ para obter uma aproximação ao valor de I um erro inferior ao pretendido. Assim,

$$I \approx \frac{1}{12}[f(0) + 4f(0{,}25) + 2f(0{,}5) + 4f(0{,}75) + f(1)] = 1{,}377903843.$$

Como só podemos garantir duas casas decimais correctas $I \approx 1{,}38$.

6.3 Problemas

6.3.1 Exercícios para resolver nas aulas

Exercício 6.13 É dada a seguinte tabela de valores de uma certa função v:

t_i	0	60	120	180	240	300
$v(t_i)$	0,0000	0,0824	0,2747	0,6502	1,3851	3,229

.

1. Determine uma aproximação para $v'(180)$ usando: i. Diferenças progressivas; ii. Diferenças regressivas; iii. Diferenças centradas.

2. Como poderia proceder para determinar uma aproximação para $v'(300)$? Justifique.

Exercício 6.14 Calcule a ordem de precisão da seguinte fórmula para a aproximação numérica

$$f'(x_i) \approx \frac{f(x_{i-2}) - 6f(x_{i-1}) + 3f(x_i) + 2f(x_{i+1})}{6h},$$

onde h é distância entre os pontos x_j, $j = i-2, ..., i+1$.

Exercício 6.15 É dada a seguinte tabela de valores de uma certa função f:

x_i	3,1	3,2	3,3	3,4	3,5
$f(x_i)$	0,0	0,6	1,0	1,2	1,3

.

1. Determine aproximações para $f'(3{,}1)$ e $f'(3{,}5)$ usando interpolação linear.
2. Determine aproximações para $f''(3{,}3)$.
3. Determine o polinómio interpolador de Hermite de f no suporte $\{3{,}1; 3{,}5\}$.

Exercício 6.16 Considere os valores de $f(x) = xe^x$ dados na seguinte tabela

x_i	1,8	1,9	2,0	2,1	2,2
$f(x_i)$	10,889365	12,703199	14,778112	17,148957	19,855030

.

Aproxime o valor de $f'(2,0)$ usando diferenças finitas e compare com o valor exato.

Exercício 6.17 A tabela a seguir mostra a distância percorrida por um veículo, em função do tempo

t (s)	0	10	20	30	40	50	60
y (m)	0	50	125	205	280	350	410

.

Seja $v(t)$ a velocidade do veículo no instante t, $v = \dfrac{dy}{dt}$. Determine uma aproximação para $v(50)$ usando diferenças regressivas, progressivas e centradas.

Exercício 6.18 A produção de citrinos em Itália foi monitorizada, desde 1970 até 1990, como é mostrado na tabela seguinte:

Ano	1970	1980	1985	1990
Produção ($\times 10^5$ Kg)	24001	25961	34336	29036

.

Determine uma aproximação para a taxa de crescimento da produção de citrinos no ano de 1990.

Exercício 6.19 A taxa de arrefecimento de um corpo pode ser expressa por

$$\frac{dT}{dt} = -k(T - T_a)$$

onde T e T_a são as temperaturas do corpo e do meio circundante (em graus Celsius), respectivamente, e k é uma constante de proporcionalidade (por minuto). Se uma esfera de metal aquecida a 90^oC é mergulhada em água mantida à temperatura constante de $T_a = 20^o$C, a temperatura da esfera toma os seguintes valores:

Tempo (min.)	0	5	10	15	20	25
Temperatura (^{o}C)	90	62,5	45,8	35,6	29,5	25,8

.

1. Use diferenciação numérica para aproximar $\frac{dT}{dt}$ em cada momento.
2. Use a alínea anterior para obter uma estimativa para a constante de proporcionalidade k.

Exercício 6.20 (Matlab) Os valores seguintes representam a evolução no tempo do número $n(t)$ de indivíduos de uma dada população.

t (meses)	0	0,5	1	1,5	2	2,5	3
$n(t)$	100	147	178	192	197	199	200

.

Utilizar estes dados para aproximar a taxa de variação desta população, usando diferentes fórmulas. Em seguida, comparar com a taxa exacta $n'(t) = 2n(t) - 0{,}01n^2(t)$.

Exercício 6.21 (Matlab) Considere-se o deslocamento de um carro numa recta. Use os dados da tabela (tempo gasto e distância percorrida) para aproximar o valor da velocidade nos instantes referidos.

Tempo (s)	0	3	5	8	10	13
Distância percorrida (m)	0	225	383	623	742	993

.

Exercício 6.22 (Matlab) Considere a função $f(x) = e^{-2x} - x$.

1. Determine o valor exacto de $f'(2)$.
2. Aproxime o valor de $f'(2)$, recorrendo a diferenças centradas, com $h = 0{,}5$, ou seja, usando os pontos $x = 2 \pm 0{,}5$. A seguir, diminua os incrementos h de $0{,}1$ até $h = 0{,}1$.
3. Repita o procedimento da alínea anterior com diferenças progressivas e regressivas.
4. Compare os valores obtidos nas duas alíneas anteriores e compare com o valor exacto da derivada.

Exercício 6.23 (Matlab) Os dados da tabela indicam a altura h em diferentes instantes dum foguetão espacial em movimento ascendente vertical. Use diferenciação numérica para completar a tabela.

Tempo (s)	0	4	8	12	16	20
Altura (km)	0	0,84	3,53	8,41	15,97	27,00
Velocidade (km/s)						

.

Exercício 6.24 Determine um valor aproximado de

$$I = \int_0^1 e^x \cos x dx,$$

com uma casa decimal correcta, usando a fórmula do ponto médio.

Exercício 6.25 Determine valores aproximados para

$$\int_0^1 e^{-x} dx,$$

usando a fórmula do ponto médio e a fórmula do trapézio. Indique um limite superior para o erro cometido em cada um dos casos.

Exercício 6.26 Considere o integral

$$I = \int_0^1 (2t^4 - 0{,}25)\ dt.$$

Use a fórmula dos trapézios para obter um valor aproximado de I com um erro absoluto inferior a $0{,}1$.

Exercício 6.27 Considere a seguinte tabela da função f

x_i	0,00	0,50	1,00
$f(x_i)$	4,76	1,05	0,00

.

1. Determine o valor aproximado do integral $\int_0^1 f(x)dx$, a partir dos dados da tabela, usando a fórmula dos trapézios composta.

2. Sabendo que
$$\max_{x\in[0,1]} |f''(x)| \leqslant 2$$
determine um majorante para o erro absoluto cometido na aproximação anterior.

Exercício 6.28 Considere os seguintes integrais

$$\int_0^1 x\ dx, \quad \int_0^1 (x^2+10x+3)\ dx, \quad \int_0^1 \cos(\pi x)\ dx.$$

1. Para quais dos integrais anteriores são as fórmulas do ponto médio e dos trapézios exactas?

2. Usando a fórmula de Simpson, determine um valor aproximado para os três integrais de tal forma que o erro absoluto cometido seja inferior a 10^{-2}.

Exercício 6.29 Considere a seguinte igualdade $\pi = 4\int_0^1 \frac{1}{1+x^2}dx$.

1. Calcule o número de pontos que deve considerar na fórmula do trapézio por forma a aproximar o valor de π com uma casa decimal correcta.

2. Determine um valor aproximado do integral usando a fórmula do trapézio e o número de pontos calculado na alínea anterior.

Exercício 6.30 Considere o integral

$$I(f) = \int_0^1 e^x\ dx.$$

1. Determine quantos pontos necessita para aproximar o valor de $I(f)$ com um erro absoluto inferior a $5\cdot 10^{-4}$, usando a regra dos trapézios e de Simpson.

2. Use a regra de Simpson com cinco pontos e determine um valor aproximado para $I(f)$.

Exercício 6.31 Seja $I = \int_0^\pi xe^{2x}dx$.

1. Qual o menor número de pontos que deve considerar na fórmula do trapézio por forma a aproximar o valor do integral com uma casa decimal correcta.

2. Calcule o valor aproximado de I de acordo com a alínea anterior.

3. Repita as alíneas anteriores usando, agora, a fórmula de Simpson.

Exercício 6.32 Seja I_1 e I_2 os valores obtidos pela fórmula composta do trapézio, aplicada com dois passos de comprimentos diferentes H_1 e H_2, ao cálculo aproximado de $I(f) = \int_a^b f(x)dx$. Verifique que, se f'' variar pouco em $]a,b[$, o valor

$$I_R = I_1 + \frac{I_1 - I_2}{(H_2/H_1)^2 - 1}$$

dá uma melhor aproximação de $I(f)$ do que I_1 e I_2. Esta técnica designa-se por método de extrapolação de Richardson.

Exercício 6.33 Considere a seguinte tabela da função $f(x)$:

x_i	0,0	0,2	0,4	0,6	0,8	1,0
$f(x_i)$	1,00	0,83	0,71	0,62	0,36	0,30

.

1. Será possível calcular um valor aproximado para o integral $I = \int_0^1 f(x)dx$, usando a fórmula de Simpson ou a regra dos trapézios, através da tabela, com um erro que não exceda 10^{-3}? Justifique a sua resposta.
2. Calcule um valor aproximado de I e indique uma estimativa para o erro cometido.

Exercício 6.34 Pretende calcular-se um valor aproximado para o integral $I = \int_1^2 \ln\frac{1}{x}dx$.

1. Use a fórmula de Simpson para obter I com 3 casas decimais correctas.
2. Sem calcular o valor exacto de I, diga, justificando, se a aproximação calculada é por defeito ou por excesso.

Exercício 6.35 Considere a seguinte equação diferencial $y'(t)+a(t)y(t) = 0$. A solução desta equação é da forma $y(t) = y(0)e^{-\int_0^t a(s)ds}$. Sabendo que $a(0) = 1$, $a(1) = 2$, $a(2) = 1$ e que $y(0) = 1$, determine uma aproximação para $y(2)$.

Exercício 6.36 A massa que é libertada por um reactor num dado período de tempo é dada por

$$M = \int_{t_1}^{t_2} QCdt,$$

onde t_1 e t_2 são os momentos inicial e terminal, respectivamente. Usando $Q = 5$ m^3/min e os dados da tabela

t (min.)	0	10	20	30	40
C (mg/m^3)	10,00	35,00	54,73	52,16	37,07

,

aproxime o valor da masssa libertada pelo reactor nos primeiros 40 minutos.

Exercício 6.37 Pretende-se determinar uma fórmula para aproximar o valor do integral

$$\int_{-1}^{1} f(x)\ dx$$

da forma

$$Af\left(-\frac{1}{\sqrt{3}}\right) + Bf\left(\frac{1}{\sqrt{3}}\right)$$

que seja exata para polinómios de grau inferior ou igual a 1. Determine os coeficientes A e B.

Exercício 6.38 Determine A e B por forma a que a fórmula dos trapézios corrigida

$$\int_0^1 f(t)\ dt \approx \frac{1}{2}(f(0)+f(1)) + Af'(0) + Bf'(1)$$

tenha o maior grau de exatidão possível.

Exercício 6.39 Determine os pesos e os nós na fórmula de integração

$$\int_0^1 f(x)dx \approx w_1 f(0) + w_2 f(x_2)$$

por forma a que tenha um grau de precisão o mais elevado possível.

Exercício 6.40 Construa uma regra de integração da forma

$$I(f) = \int_{-1}^1 f(x)dx \approx A_0 f(-\frac{1}{2}) + A_1 f(0) + A_2 f(\frac{1}{2})$$

de modo a ter grau de exactidão igual a 2.

Exercício 6.41 (Matlab) Determine uma aproximação do valor do integral

$$\int_0^1 \frac{\sin x}{x}\ dx,$$

usando uma regra de integração apropriada.

Exercício 6.42 (Matlab) Um carro de corrida completa uma volta num circuito em 84 s. A velocidade do carro (em metros por segundo) em diferentes instantes temporais (em segundos) é dada na seguinte tabela.

Tempo	0	6	12	18	24	30	36	42	48	54	60	66	72	78	84
Velocidade	124	134	148	156	147	133	121	109	99	85	78	89	104	116	123

Determine um valor aproximado do comprimento do circuito.

Exercício 6.43 (Matlab) Determine um valor aproximado do integral

$$\int_0^\pi e^x \cos x\ dx$$

usando a fórmula dos trapézios composta, com 2, 20, 200 e 2000 subintervalos.

Exercício 6.44 (Matlab) Determine o comprimento aproximado do arco do gráfico da função $f(x) = x^3 - x$, entre os pontos (-1,0) e (2,6), usando a fórmula do trapézio composta, com 4 subintervalos.

Exercício 6.45 (Matlab) Considere a função $f(x) = e^x + 2x$.

1. Calcule uma aproximação para a raiz de $f(x)$ aplicando o método de Newton 2 vezes.
2. Utilizando a fórmula de Simpson, aproxime a área da região limitada por $y \leq e^x$, $y \geq -2x$ e $x \leq 0$.

Exercício 6.46 (Matlab) Determine o número mínimo de subintervalos para aproximar, usando a fórmula composta do ponto médio com erro inferior a 10^{-4}, os integrais das seguintes funções: $f_1(x) = 1/(1+(x-\pi)^5)$, em [0,5], $f_2(x) = e^x \cos x$, em $[0, \pi]$ e $f_3(x) = \sqrt{x(1-x)}$, em [0,1].

Exercício 6.47 (Matlab) Consideremos um condutor eléctrico esférico de raio arbitrário r e condutividade σ. Pretendemos calcular a distribuição da densidade de corrente **j** em função de r e t (tempo), conhecendo a distribuição inicial da densidade de corrente $\rho(r)$. O problema pode ser resolvido usando as relações entre a densidade de corrente, o campo eléctrico e a densidade de carga e observando que, pela simetria da configuração, $\mathbf{j}(r,t) = j(r,t)\mathbf{r}/|\mathbf{r}|$, em que $j = |\mathbf{j}|$. Obtém-se

$$j(r,t) = \gamma(r)e^{-\sigma t/\varepsilon_0}, \ \gamma(r) = \frac{\sigma}{\varepsilon_0 r^2}\int_0^r \rho(\xi)\xi^2 \, d\xi,$$

onde $\varepsilon_0 = 8{,}859 \times 10^{-12}$ farad/m é a constante dieléctrica do vazio. Usando a fórmula de Simpson composta, determine a função $\gamma(r)$, para $r = k/10$ m com k=1,...,10, $\rho(\xi) = e^{\xi}$ e $\sigma = 0{,}36$ W/(mK). (Recorde que: m=metros, W=watts, K=graus Kelvin).

Exercício 6.48 (Matlab) A fim de planificar uma sala para raios infravermelhos, estamos interessados em calcular a energia emitida por um corpo negro (isto é, um objecto capaz de irradiar em todo o espectro à temperatura ambiente) no espectro (infravermelho) compreendido entre os comprimentos de onda 3 μm e 14 μm. A solução deste problema obtém-se calculando o integral

$$E(T) = 2{,}39 \times 10^{-11} \int_{3\times 10^{-4}}^{14\times 10^{-4}} \frac{dx}{x^5\left(e^{1{,}432/(Tx)-1}\right)},$$

que é a equação de Planck para a energia $E(T)$, onde x é o comprimento de onda (em cm) e T a temperatura (em Kelvin) do corpo negro. Recorra à fórmula de Simpson adaptativa para determinar a função $E(T)$, com $T = 213$ K.

6.3.2 Exercícios de aplicação à engenharia

Exercício 6.49 Num circuito eléctrico com voltagem aplicada $E(t)$ e inductância L, a primeira Lei de Kirchoff dá-nos a relação

$$E(t) = LI'(t)RI(t),$$

onde R é a resistência no circuito e $I(t)$ a corrente no instante t. Suponhamos que medimos a corrente para vários valores de $t = t_i$, $i = 1, \ldots, 5$, obtendo

t_i	1,00	1,01	1,02	1,03	1,04
$I(t_i)$	3,10	3,12	3,14	3,18	3,24

,

onde tempo é medido em segundos, a corrente em amperes, a inductância é uma constante dada por $L = 0{,}98$ henries e a resistência é $0{,}142$ ohms. Aproxime a voltagem E nos valores de t dados na tabela.

Exercício 6.50 Os valores seguintes representam a evolução no tempo do número $N(t)$ de indivíduos de uma dada população cuja taxa de crescimento é constante ($b = 2$) e cuja taxa de mortalidade é $d(t) = 0{,}01N(t)$:

t (meses)	0	0,5	1	1,5	2	2,5	3
N	100	147	178	192	197	199	200

.

1. Utilize os dados da tabela para aproximar com a maior precisão possível a taxa de variação desta população.

2. Compare os resultados obtidos na alínea anterior com a taxa exacta $N'(t) = 2N(t) - 0{,}01N^2(t)$.

Exercício 6.51 A altura $q(t)$ atingida no tempo t por um fluido contido num reservatório cilíndrico rectilíneo de raio $R = 1$ m tendo na sua base um orifício circular de raio $r = 0{,}1$ m, foi medida em cada 5 segundos, tendo-se registado os seguintes valores:

t	0	5	10	15	20
$q(t)$	0,6350	0,5336	0,4410	0,3572	0,2822

.

1. Utilize os dados da tabela para aproximar com a maior precisão possível a velocidade de esvaziamento $q'(t)$.

2. Compare os resultados obtidos na alínea anterior com velocidade prevista pela lei de Torricelli: $q'(t) = -\gamma(r/R)^2\sqrt{2gq(t)}$, onde g é a aceleração da gravidade e $\gamma = 0{,}6$ é um factor de correcção.

Exercício 6.52 Fugacidade é o termo usado na engenharia para descrever a trabalho resultante de um processo isotérmico. Para um gás ideal, a fugacidade f é igual à pressão P, mas para os gases reais,

$$\ln \frac{f}{P} = \int_0^P \frac{C-1}{P} dp$$

onde C é um factor de compressibilidade determinado experimentalmente. Para o metano os valores de C são:

P (atm.)	C
1	0,9940
10	0,9370
20	0,8683
40	0,7043
60	0,4515

P (atm.)	C
80	0,3429
120	0,4259
160	0,5252
250	0,7468
400	1,0980

.

Escreva um programa que calcule o valor de f correspondente a cada valor da pressão dado na tabela. Assuma que o valor de C varia linearmente entre os valores calculados e que C tende para um quando P tende para zero.

Exercício 6.53 A função

$$f(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

é usada com muita frequência em disciplinas tão diversas como a teoria das probabilidades, distribuição de calor, difusão de matérias, etc. Usando uma das regras de integração estudadas, calcule uma aproximação para o valor do referido integral indicando um majorante para o erro cometido.

Exercício 6.54 Uma partícula de massa m movendo-se num fluído está sujeita a uma resistência de viscosidade R, que é função da velocidade v. A relação entre a resistência R, a velocidade v e o tempo t é dada pela equação

$$t = \int_{v(t_0)}^{v(t)} \frac{m}{R(u)} du.$$

onde R é a resistência no circuito e $I(t)$ a corrente. Suponhamos que $R(v) = -v\sqrt{v}$ para um fluído particular, onde R é dado em newtons e v em metros/segundo. Se $m = 10$ kg e $v(0) = 10$ m/seg aproxime o tempo necessário para a partícula reduzir a sua velocidade para $v = 5$ m/seg.

Exercício 6.55 A intensidade de luz com comprimento de onda λ viajando através de uma grelha de difracção com n aberturas a um ângulo θ é dada por

$$I(\theta) = \frac{n^2}{k} \sin^2 k,$$

onde

$$k = \frac{\pi n d \sin\theta}{\lambda}$$

e d é a distância entre cada abertura. Um laser de hélio-néon com comprimento de onda $\lambda = 632{,}8 \times 10^{-9}$ m emite uma banda estreita de luz, dada por $-10^{-6} < \theta < 10^{-6}$, através de uma grelha com 10000 aberturas separadas por 10^{-4} m. Obtenha um valor aproximado para a intensidade de luz total que sai da grelha

$$\int_{-10^{-6}}^{10^{-6}} I(\theta)\, d\theta.$$

Capítulo 7

Equações diferenciais ordinárias

As primeiras equações diferenciais são tão antigas quanto o cálculo diferencial. Newton considerou-as, em 1671, no seu tratado de cálculo diferencial e discutiu a sua solução por integração e por expansão em série. Leibniz, o segundo inventor do cálculo, chegou às equações diferenciais por volta de 1676 considerando o *problema geométrico do inverso das tangentes*: para que curva $y(x)$ a tangente em cada ponto P tem um comprimento constante (com o eixo dos x's), digamos a? Este problema conduziu à equação $y' = -y/\sqrt{a^2 - y^2}$.

Em 1696, Johann Bernoulli (1667-1748) convidou os mais ilustres matemáticos do seu tempo para resolver o *problema da braquistócrona* (curva de tempo mínimo), principalmente para refutar a resposta, que esperava errada, do seu irmão Jacob Bernoulli (1657-1705). O problema consistia em determinar a curva $y(x)$ que une dois pontos P_0 e P_1 de tal modo que um ponto, partindo de P_0 e "deslizando", nessa curva, sujeito apenas a forças gravíticas, atinja P_1 no menor tempo possível. A resposta a este problema foi dada dada por vários matemáticos (inclusivé Jacob Bernoulli) e é, como se sabe, a *ciclóide*. Essa curva pode ser determinada como sendo a solução de uma equação diferencial ordinária.

Muitos problemas da engenharia e da ciência têm como modelo equações diferenciais. Neste curso iremos efectuar uma breve introdução ao estudo dos métodos numéricos para a resolução de problemas que envolvem equações diferenciais. Os problemas que iremos considerar serão de dois tipos: problemas com condição inicial e problemas com condições de fronteira.

7.1 O problema de Cauchy

Consideremos uma equação diferencial ordinária de primeira ordem, isto é, uma equação da forma

$$y'(t) = f(t, y(t)), \qquad t \in [t_0, T], \tag{7.1}$$

em que $f : [t_0, T] \times \mathbb{R} \longrightarrow \mathbb{R}$. O estudo que iremos efectuar para este tipo de equações pode ser facilmente generalizado a sistemas de equações diferencias ordinárias de primeira ordem, isto é, para o caso em que $f : [t_0, T] \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$. Por uma questão de simplificação de exposição optámos por apresentar o estudo para o caso escalar ($N = 1$).

Antes de se pensar em resolver uma determinada equação diferencial há que garantir que essa equação tem solução e que é única. Note-se que a solução equação (7.1), se existir, não é única pois, ao integrarmos, introduzimos sempre uma constante de integração.

Uma das condições para obter a unicidade da solução consiste em especificar $y(t)$ num ponto qualquer do intervalo $[t_0, T]$, usualmente o ponto inicial t_0. Ficamos assim com o

problema de condição inicial (PCI)

$$\begin{cases} y'(t) &= f(t, y(t)), \qquad t \in]t_0, T] \\ y(t_0) &= y_0 \end{cases}, \tag{7.2}$$

também chamado problema de Cauchy, em homenagem a Augustin-Louis Cauchy (1789-1857).

Apesar de contornado este problema ainda não temos a garantia da existência e unicidade da solução do PCI (7.2). Antes de apresentarmos o teorema que estabelece as condições suficientes para que o problema tenha solução única consideremos a definição seguinte devida a Rudolf Otto Sigismund Lipschitz (1832-1903).

Definição 7.1 (Função lipschitziana) *Uma função $f(t, y)$ verifica a* condição de Lipschitz *(ou é* lipschitziana*), na variável y, num conjunto $D \subset \mathbb{R}^2$ se existir uma constante $L > 0$ tal que*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|,$$

sempre que $(t, y_1), (t, y_2) \in D$. A L chama-se constante de Lipschitz.

Exercício 7.1 Prove que a função $f(t, y) = t|y|$ é lipschitziana, na variável y, no conjunto

$$D = \{(t, y) \in \mathbb{R}^2 : 1 \leq t \leq 2;\ -3 \leq y \leq 4\}.$$

Resolução: Temos que

$$|f(t, y_1) - f(t, y_2)| = |t|y_1| - t|y_2|| \leq 2||y_1| - |y_2|| \leq 2|y_1 - y_2|.$$

Logo, a constante de Lipschitz é $L = 2$.

O teorema seguinte, atribuído a Charles Émile Picard (1856 -1941), estabelece condições suficientes para que um problema com condição inicial tenha solução única.

Teorema 7.1 (Picard) *Seja $f(t, y)$ uma função contínua (nas variáveis t e y) e lipschitziana (na variável y) em $D = \{(t, y) : t_0 \leq t \leq T, y \in \mathbb{R}\}$. Então o PCI (7.2) tem solução única $y(t) \in C^1([t_0, T])$.*

A demonstração deste teorema estabelece um processo iterativo de aproximação da solução do PCI (7.2) conhecido por método de Picard. Se f for contínua em relação a t, determinar a solução do PCI (7.2) é equivalente a determinar y, continuamente diferenciável, que verifica

$$y(t) = y_0 + \int_a^t f(\tau, y(\tau))d\tau. \tag{7.3}$$

O que se prova na demonstração do Teorema de Picard é que a sucessão de funções $\{u_j(t)\}$, definida recursivamente por

$$\begin{aligned} u_0(t) &= y_0, \\ u_{j+1}(t) &= y_0 + \textstyle\int_a^t f(\tau, u_j(\tau))d\tau. \quad j = 0, 1, \ldots, \end{aligned}$$

converge para a única solução de (7.3).

Como corolário do Teorema de Picard temos o seguinte resultado que apresentamos, igualmente, sem demonstração.

Corolário 7.2 *Suponhamos que $f(t,y)$ está definida num conjunto convexo $D \subset \mathbb{R}^2$, isto é, num conjunto $D \subseteq R^2$ tal que, para qualquer $(t_1,y_1),(t_2,y_2) \in D$, se verifica*

$$((1-\theta)t_1+\theta t_2,(1-\theta)y_1+\theta y_2) \in D, \qquad \theta \in [0,1].$$

Se existir uma constante $L>0$ tal que

$$\left|\frac{\partial f}{\partial y}(t,y)\right| \leq L, \qquad \forall (t,y) \in D,$$

então f satisfaz a condição de Lipschitz, na variável y, com L a respectiva constante e, como tal, o PCI (7.2) tem solução única $y(t) \in C^1([t_0,T])$.

Note-se que o conjunto $D=\{(t,y): t_0 \leq t \leq T, y \in \mathbb{R}\}$ é, obviamente, convexo.

Exercício 7.2 Mostre que o problema de condição inicial

$$\begin{cases} y'(t) &= \dfrac{1}{1+y^2}, \qquad t \in]a,b] \\ y(a) &= 0 \end{cases}$$

tem solução única.

Resolução: Seja $D=\{(t,y): a \leq t \leq b, y \in \mathbb{R}\}$ e

$$f(y)=\frac{1}{1+y^2}.$$

Vamos provar que a função

$$\left|\frac{\partial f}{\partial y}(t,y)\right| = \left|\frac{-2y}{(1+y^2)^2}\right|$$

é limitada em D. Para isso há que determinar

$$L=\max_{y\in\mathbb{R}}\left|\frac{2y}{(1+y^2)^2}\right|.$$

Como a função que queremos provar limitada é par temos que

$$L=\max_{y\in\mathbb{R}_0^+}\frac{2y}{(1+y^2)^2}.$$

Consideremos

$$g(y)=\frac{2y}{(1+y^2)^2}.$$

Como

$$g'(y)=0 \Rightarrow y=\pm\frac{\sqrt{3}}{3}$$

temos que

$$L=\max\left\{g(0),\ g\left(\frac{\sqrt{3}}{3}\right),\ \lim_{y\to+\infty}g(y)\right\}=\max\{0,\ 0{,}6594,\ 0\}=0{,}6594.$$

Está assim provado o pretendido.

7.2 Métodos numéricos para o problema de Cauchy

Consideremos, de novo, o PCI (7.2) verificando as condições do Teorema de Picard. Os métodos numéricos que iremos considerar para resolver este problema são métodos discretos, isto é, são métodos que determinam aproximações $u_0, u_1, \ldots, u_n$ para a solução exacta $y_0 = y(t_0), y_1 = y(t_1), \ldots, y_n = y(t_n)$ nos pontos distintos da malha

$$t_0 < t_1 < \cdots < t_{n-1} < t_n = T.$$

Às distâncias $h_i = t_i - t_{i-1}$, $i = 1, \ldots, n$, dá-se o nome de passos de discretização (ou medidas do passo) da malha. Se os passos forem todos iguais a malha diz-se uniforme ou de passo constante. Caso contrário diz-se de passo variável. Neste curso vamos apenas considerar malhas uniformes, isto é, tais que $t_i = t_0 + ih$, $i = 0, \ldots, n$, onde $h = \frac{T-t_0}{n}$.

Os métodos numéricos permitem determinar valores $u_i \approx y_i = y(t_i)$ por meio de relações de recorrência deduzidas do PCI (7.2) de modo a que o valor de u_{i+1} venha expresso em função de $u_i, u_{i-1}, \ldots, u_0$, sendo $u_0 = y_0$. A $\{u_0 = y_0, u_1, u_2, \ldots, u_{n-1}, u_n\}$ chama-se solução numérica. É usual agrupar os métodos numéricos para a resolução de problemas de condição inicial em duas grandes classes.

- Métodos de passo único. São métodos que determinam o valor de u_{i+1} apenas à custa de u_i.
- Métodos de passo múltiplo. São métodos que determinam o valor de u_{i+1} à custa de $u_i, u_{i-1}, \ldots, u_{i-r+1}$. Neste caso diz-se que o método é de r passos.

Neste curso iremos apenas abordar os métodos de passo único. Estes métodos, por sua vez, podem ainda ser de dois tipos.

- Métodos explícitos. São métodos em que o valor de u_{i+1} é determinado directamente a partir de u_i. Estes métodos podem ser escritos na forma

$$u_{i+1} = u_i + h\phi(t_i, u_i; h). \tag{7.4}$$

- Métodos implícitos. São métodos em que o valor de u_{i+1} depende implicitamente de si mesmo através de f. Estes métodos podem ser escritos na forma

$$u_{i+1} = u_i + h\phi(t_i, t_{i+1}, u_i, u_{i+1}; h). \tag{7.5}$$

A função ϕ que define os métodos (7.4) e (7.5) é chamada função de iteração ou função incremento do método numérico.

7.2.1 Métodos baseados na série de Taylor

Consideremos o PCI (7.2) com f uma função suficientemente diferenciável nas variáveis t e y. Então, fazendo o desenvolvimento em série de Taylor temos

$$y(t) = y(t_0) + (t - t_0)y'(t_0) + \frac{(t-t_0)^2}{2!}y''(t_0) + \cdots.$$

As derivadas que aparecem nesta expressão não são conhecidas explicitamente visto que a solução também não é conhecida. No entanto, podemos escrever

$$\begin{aligned} y'(t) &= f(t,y), \\ y''(t) &= \frac{df}{dt}(t,y) = (f_t + f_y y')(t,y) = (f_t + f_y f)(t,y), \\ y'''(t) &= \frac{d^2 f}{dt^2}(t,y) = (f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_t f_y + f_y^2 f)(t,y), \\ &\vdots \end{aligned}$$

onde

$$f_t(t,y) = \frac{\partial f}{\partial t}(t,y), \qquad f_y(t,y) = \frac{\partial f}{\partial y}(t,y), \qquad \ldots.$$

Por razões práticas temos que limitar o número de termos na expansão em série de $y(t)$ a um número razoável, o que nos conduz a restrições nos valores de t para os quais a expansão nos dá uma boa aproximação.

Se tomarmos a série de Taylor truncada temos, para $t = t_1$,

$$y(t_1) \approx u_1 = u_0 + hf(t_0,u_0) + \frac{h^2}{2}f'(t_0,u_0) + \cdots + \frac{h^k}{k!}f^{(k-1)}(t_0,u_0),$$

onde

$$f^{(j)}(t_0,u_0) = \frac{d^j f}{dt^j}(t_0,u_0).$$

Podemos definir assim, para cada $k = 1, 2 \ldots$, um método de passo único explícito que permite obter soluções aproximadas $u_i \approx y(t_i)$ da forma (7.4) em que

$$\phi(t,u;h) = f(t,u) + \frac{h}{2}f'(t,u) + \cdots + \frac{h^k}{k!}f^{(k-1)}(t,u). \tag{7.6}$$

Os métodos assim definidos são conhecidos por métodos de Taylor. O método desta classe mais simples é quando $k = 1$, isto é, o método

$$u_{i+1} = u_i + hf(t_i,u_i), \quad i = 0,\ldots,n, \qquad u_0 = y_0, \tag{7.7}$$

designado por método de Euler progressivo (ou explícito).

O seguinte algoritmo permite determinar a solução do PCI (7.2) em $t = T$, usando o método com função incremento (7.6).

Algoritmo 7.1 Método de Taylor

Dados: n, k, t_0, T e y_0

$h := (T - t_0)/n$

$t := t_0$

$u := y_0$

Para i de 1 até n fazer

$\phi := 0$

Para j de 1 até k fazer

$\phi := \phi + f^{(j)}(t, u)h^j/j!$

$u := u + h\phi$

$t := t + h$

Resultado: $y(T) \approx u$

Exercício 7.3 Considere o problema de condição inicial

$$\begin{cases} y'(t) &= -2y \\ y(0) &= 1 \end{cases}.$$

Determine, usando o método de Euler progressivo, o valor aproximado de $y(1)$, fazendo $h = 1$, $h = 0.5$ e $h = 0.25$. Compare os resultados obtidos sabendo que $y(t) = e^{-2t}$.

Resolução: A solução exacta deste problema é $y(1) = 0{,}135335283$. Consideremos agora as soluções numéricas para os três casos propostos. Seja $f(y) = -2y$.

- $h = 1$

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(1) \approx u_1 &= u_0 + hf(u_0) = 1 + 1 \times (-2) = -1. \end{aligned}$$

  Logo $|y(1) - u_1| = 1{,}135335283$.

- $h = 0.5$

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(0{,}5) \approx u_1 &= u_0 + hf(u_0) = 1 + 0{,}5 \times (-2) = 0 \\ y(1) \approx u_2 &= u_1 + hf(u_1) = 0 + 0{,}5 \times 0 = 0. \end{aligned}$$

  Logo $|y(1) - u_2| = 0{,}135335283$.

- $h = 0{,}25$

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(0{,}25) \approx u_1 &= u_0 + hf(u_0) = 1 + 0{,}25 \times (-2) = 0{,}5 \\ y(0{,}5) \approx u_2 &= u_1 + hf(u_1) = 0{,}5 + 0{,}25 \times (-1) = 0{,}25 \\ y(0{,}75) \approx u_3 &= u_2 + hf(u_2) = 0{,}25 + 0{,}25 \times (-0{,}5) = 0{,}125 \\ y(1) \approx u_4 &= u_3 + hf(u_3) = 0{,}125 + 0{,}25 \times (-0{,}25) = 0{,}0625. \end{aligned}$$

  Logo $|y(1) - u_4| = 0{,}072835283$.

Nota-se que, quanto menor for a medida do passo mais pequeno é o erro cometido.

Exercício 7.4 Seja dado o problema de condição inicial

$$\begin{cases} y'(t) &= \dfrac{1}{1+y^2} \\ y(0) &= 1 \end{cases}.$$

Use o método de Taylor, com $k = 2$, para determinar o valor aproximado de $y(1)$, fazendo $h = 0{,}5$.

Resolução: Seja $f(y) = (1+y^2)^{-1}$. Temos que o método de Taylor com $k = 2$ é dado por

$$u_{i+1} = u_i + hf(u_i) + \frac{h^2}{2}\frac{df}{dt}(u_i) = u_i + h\frac{1}{1+u_i^2} - h^2\frac{u_i}{(1+u_i^2)^3}.$$

Assim, fazendo $h = 0{,}5$ temos

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(0{,}5) \approx u_1 &= 1 + 0{,}5 \times \frac{1}{2} - 0{,}25 \times \frac{1}{8} = 1{,}21875 \\ y(1) \approx u_2 &= 1{,}21875 + 0{,}5 \times \frac{1}{2{,}485351563} - 0{,}25 \times \frac{1{,}21875}{15{,}35194798} = 1{,}4. \end{aligned}$$

7.2.2 Métodos de passo único implícitos

Os métodos de passo único implícitos da forma (7.5) também têm muita relevância prática. Não havendo possibilidade de explicitar o valor de u_{i+1} temos necessidade de o calcular resolvendo a equação (geralmente não linear)

$$u_{i+1} - u_i - h\phi(t_i, t_{i+1}, u_i, u_{i+1}; h) = 0.$$

Usualmente considera-se um método numérico na resolução desta equação.

Se considerarmos o método de Newton, a primeira questão a resolver é a da determinação de uma aproximação inicial $u_{i+1}^{(0)}$. Normalmente toma-se para aproximação inicial o valor de u_i; outra hipótese será a de considerar a aproximação inicial obtida pela aplicação de um método explícito. Deteminado o valor de $u_{i+1}^{(0)}$ temos que

$$u_{i+1}^{(k+1)} = u_{i+1}^{(k)} - \frac{g(u_{i+1}^{(k)})}{g'(u_{i+1}^{(k)})}, \qquad k = 0, 1, \ldots,$$

sendo

$$g(u) = u - u_i - h\phi(t_i, t_{i+1}, u_i, u; h).$$

Os métodos implícitos são usados visto que, em geral, são mais precisos e menos sensíveis a erros que os métodos explícitos. Por outro lado, o esforço computacional exigido no cálculo de u_{i+1} é, para os métodos implícitos, muito maior. Assim, estes métodos só devem ser usados quando há necessidade de uma precisão muito elevada em problemas sensíveis a erros.

Exemplos comuns de métodos implícitos são o chamado método de Euler regressivo (ou implícito), dado pela expressão

$$u_{i+1} = u_i + hf(t_{i+1}, u_{i+1}), \quad i = 0, \ldots, n-1, \qquad u_0 = y_0,$$

e o método dos trapézios ou método de Crank-Nicolson, em homenagem a John Crank (1916-2006) e Phyllis Nicolson (1917-1968), dado por

$$u_{i+1} = u_i + \frac{h}{2}(f(t_i, u_i) + f(t_{i+1}, u_{i+1})), \quad i = 0, \ldots, n-1, \qquad u_0 = y_0.$$

Exercício 7.5 Considere o problema de condição inicial

$$\begin{cases} y' &= -30y \\ y(0) &= 1 \end{cases}$$

e os método de Euler progressivo e regressivo. Usando cada um dos métodos determine a solução do problema em $t = 1$ com $h < 1$, comparando os resultados obtidos.

Resolução: Seja $f(y) = -30y$ e consideremos $h = 0{,}5$. Vamos aplicar os dois métodos separadamente.

1. Método de Euler progressivo

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(0{,}5) \approx u_1 &= 1 + 0.5 \times (-30) = -14 \\ y(1) \approx u_2 &= -14 + 0{,}5 \times (-30 \times (-14)) = 196. \end{aligned}$$

2. Método de Euler regressivo

$$\begin{aligned} y(0) = u_0 &= y_0 = 1 \\ y(0{,}5) \approx u_1 &= 1 + 0{,}5 \times (-30u_1) = 1 - 15u_1. \end{aligned}$$

Resolvendo a equação temos que $y(0{,}5) \approx u_1 = 0{,}0625$. Continuando temos

$$y(1) \approx u_2 = 0{,}0625 + 0.5 \times (-30u_2) = 0{,}0625 - 15u_2,$$

e assim, $y(1) \approx u_2 = 3{,}9 \times 10^{-3}$.

Atendendo a que a solução exacta é dada por $y(t) = e^{-30t}$ temos que $y(1) = 9{,}36 \times 10^{-14}$.

Note-se que, enquanto o método implícito se aproxima da solução o método explícito dá um resultado completamente disparatado. Os problemas que não podem ser resolvidos por métodos explícitos são chamados *stiff* e ocorrem com muita frequência em problemas de engenharia química.

7.3 Estudo do erro

Quando se determinam valores numéricos para aproximar quantidades desconhecidas, temos necessidade de conhecer estimativas para o erro que se comete nessas aproximações. No caso dos métodos numéricos para a resolução de equações diferenciais vamos considerar dois tipos de erros: o erro de truncatura local e o erro global (ou da aproximação).

Comecemos por considerar métodos numéricos de passo único explícitos da forma

$$u_{i+1} = u_i + h\phi(t_i, u_i; h), \quad i = 0, \ldots, n-1, \qquad u_0 = y_0. \tag{7.8}$$

Pretendemos estudar o comportamento do seu erro global.

Definição 7.2 (Erro global e convergência) *Considere-se o PCI (7.2) e um método numérico de passo único explícito (7.8) que determine aproximações u_i para a solução exacta $y(t_i)$, $i = 0, 1, \ldots, n$. A $e(t_i) = y(t_i) - u_i$ chama-se* erro global *do método no ponto t_i. Se*

$$\lim_{h\to 0} \max_{1\leq i\leq n} |e(t_i)| = 0,$$

o método diz-se convergente. *O método diz-se de* ordem (de convergência) $p > 0$ *se existir um $C > 0$ tal que*

$$|e(t_i)| \leq Ch^p, \qquad i = 1, \ldots, n$$

ou, noutra notação, se $|e(t_i)| = \mathcal{O}(h^p)$, $i = 1, \ldots, n$.

O estudo do erro global pode ser difícil. Para isso, considere-se,

$$e(t_i) = y(t_i) - u_i = (y(t_i) - u_i^*) + (u_i^* - u_i),$$

com

$$u_i^* = y(t_{i-1}) + h\phi(t_{i-1}, y(t_{i-1}); h), \quad i = 1, \ldots, n, \qquad u_0 = y_0, \tag{7.9}$$

isto é, a solução numérica calculada no nível temporal t_i, supondo $u_j = y(t_j)$, $j = 0, ..., i-1$.

Definição 7.3 (Erro de truncatura e consistência) *Considere-se o PCI (7.2), verificando as condições do Teorema de Picard, e um método numérico de passo único (7.8) que determine aproximações u_i para a solução exacta $y(t_i)$, $i = 0, 1, \ldots, n$. O* erro de truncatura local *do método no ponto t_i é definido por*

$$T_i(h) = \frac{y(t_i) - u_i^*}{h}.$$

Se

$$\lim_{h\to 0} \max_{1\leq i\leq n} |T_i(h)| = 0,$$

o método diz-se consistente *com o PCI (7.2). O método diz-se de* ordem (de consistência) $p > 0$ *se existir um $C > 0$ tal que*

$$|T_i(h)| \leq Ch^p, \qquad i = 1, \ldots, n$$

ou, noutra notação, se $|T_i(h)| = \mathcal{O}(h^p)$, $i = 1, \ldots, n$.

Da definição anterior conclui-se que o erro de truncatura local é definido com sendo

$$hT_i(h) = y(t_i) - u_i^* = y(t_i) - y(t_{i-1}) - h\phi(t_{i-1}, y(t_{i-1}); h), \qquad i = 1, \ldots, n.$$

Assim, o erro local pode ser determinado através dos seguintes passos: (i) substituir na expressão que define o método numérico a solução aproximada no ponto t_i, u_i, pela solução exacta $y(t_i)$; (ii) considerar a hipótese $u_{i-1} = y(t_{i-1})$; (iii) efectuar o desenvolvimento em série de Taylor de $y(t_i)$ em torno de t_{i-1}.

Exercício 7.6 Mostre que:

1. um método é consistente se tiver, pelo menos, ordem um ou, o que é equivalente, se $\phi(t, y; 0) = f(t, y)$;
2. o erro local para o método de Taylor de função incremento (7.6) é dado por

$$T_i(h) = \frac{h^k}{(k+1)!} y^{(k+1)}(\xi), \qquad \xi \in]t_{i-1}, t_i[,$$

   ou seja, $T_i(h) = \mathcal{O}(h^k)$ e, como tal, o método tem ordem k;
3. para o método de Euler progressivo se tem $T_i(h) = \mathcal{O}(h)$, ou seja, o método tem ordem um.

O próximo teorema estabelece as condições para as quais se pode concluir que o erro global tem a mesma ordem que o erro local.

Teorema 7.3 *Seja $y(t)$ a única solução do PCI (7.2), verificando as condições do Teorema de Picard, e (7.8) um método numérico que supomos ser consistente com o problema e ter ordem p, isto é, $|T_i(h)| \leq Ch^p$, $i = 1, \ldots, n$, $p \geq 1$. Se existir $h_0 > 0$ tal que $\phi(t, y; h)$ é contínua, nas variáveis t e y, e lipschitziana, na variável y, no conjunto*

$$D = \{(t, y; h) : t_0 \leq t \leq T, y \in \mathbb{R}, 0 \leq h \leq h_0\},$$

então

$$|e(t_i)| \leq \frac{C}{L} h^p \left[e^{L(t_i - t_0)} - 1 \right], \qquad i = 1, \ldots, n.$$

sendo L a constante de Lipschitz de ϕ.

Demonstração: Considerando a definição de erro global temos que

$$e(t_i) = e(t_{i-1}) + h\left[\phi(t_{i-1}, y(t_{i-1}); h) - \phi(t_{i-1}, u_{i-1}; h)\right] + hT_i(h), \qquad i = 1, \ldots, n.$$

Uma vez que a função ϕ é lipschitziana, na variável y, e o método tem ordem $p \geq 1$ é possível concluir que

$$|e(t_i)| \leq (1 + hL)|e(t_{i-1})| + Ch^{p+1}.$$

Como $e(t_0) = 0$ obtém-se

$$|e(t_i)| \leq Ch^{p+1} \sum_{j=0}^{i-1} (1 + hL)^j = Ch^{p+1} \frac{1 - (1 + hL)^i}{1 - (1 + hL)} \leq \frac{C}{L} h^p \left[e^{ihL} - 1 \right].$$

O teorema fica assim demonstrado uma vez que $t_i = t_0 + ih$. $\square$

Note-se que a consistência, por si só, não implica convergência uma vez que existem mais tipos de erros que podem ocorrer para além do erro de truncatura local. De facto, nem as condições iniciais nem a aritmética usada estão isentas de erros. Temos portanto necessidade de garantir que os métodos usados sejam estáveis no sentido de que pequenas alterações nas condições iniciais não produzam, por aplicação do método, grandes alterações nos resultados. No caso dos métodos de passo único, o teorema anterior permite-nos estabelecer o seguinte resultado.

Corolário 7.4 *Suponhamos que o PCI (7.2) é aproximado pelo método (7.8). Se existir $h_0 > 0$ tal que $\phi(t, y; h)$ é contínua, nas variáveis t e y, e lipschitziana, na variável y, no conjunto*

$$D = \{(t, y; h) : t_0 \leq t \leq T, y \in \mathbb{R}, 0 \leq h \leq h_0\},$$

então o método (7.8): (i) é estável; (ii) é convergente se e só se é consistente.

Apesar do estudo da consistência e convergência de um método iterativo ter sido efectuado apenas para métodos explícitos, estes conceitos ainda são válidos para métodos implícitos. Para o método implícito (7.5) o erro de truncatura local é definido por

$$hT_i(h) = y(t_i) - u_i^* = y(t_i) - y(t_{i-1}) - h\phi(t_{i-1}, t_i, y(t_{i-1}), y(t_i); h), \qquad i = 1, \ldots, n.$$

Exercício 7.7 Considere o método dos trapézios na resolução de um problema de condição inicial.

1. Determine a ordem e o erro de truncatura local do método.
2. Aplique o método ao problema de condição inicial

$$\begin{cases} y'(t) &= -ty^2, \qquad t \in]0, 1] \\ y(0) &= 2 \end{cases}$$

e obtenha uma aproximação em $t = 1$ usando $h < 1$. (Considere a solução exacta positiva em $[0, 1]$.)

Resolução: 1. Atendendo à definição de erro local temos que $hT_i(h) = y(t_i) - u_i^*$. Desenvolvendo $y(t_i)$ e u_i^* em série de Taylor em torno do ponto t_{i-1}, temos que

$$y(t_i) = y(t_{i-1}) + hy'(t_{i-1}) + \frac{h^2}{2}y''(t_{i-1}) + \frac{h^3}{6}y'''(t_{i-1}) + \cdots$$

e

$$u_i^* = y(t_{i-1}) + \frac{h}{2}\left(y'(t_{i-1}) + y'(t_{i-1}) + hy''(t_{i-1}) + \frac{h^2}{2}y'''(t_{i-1}) + \cdots\right).$$

Subtraindo membro a membro vem

$$hT_i(h) = -\frac{h^3}{12}y'''(t_{i-1}) + \cdots.$$

Assim sai que

$$T_i(h) = -\frac{h^2}{12}y'''(\xi), \qquad \xi \in]t_{i-1}, t_i[.$$

Como $T_i(h) = \mathcal{O}(h^2)$ temos que o método dos trapézios tem ordem 2.

2. Seja $f(t, y) = -ty^2$ e $h = 0.5$. Assim,

$$\begin{aligned} y(0) = u_0 &= y_0 = 2 \\ y(0{,}5) \approx u_1 &= u_0 + \tfrac{h}{2}(f(t_0, u_0) + f(t_1, u_1)) = 2 - 0{,}125u_1^2. \end{aligned}$$

Vamos agora resolver a equação $0{,}125u_1^2 + u_1 - 2 = 0$. Esta equação resolve-se sem dificuldade pois

$$0{,}125u_1^2 + u_1 - 2 = 0 \Rightarrow u_1 = -9{,}6598 \text{ ou } u_1 = 1{,}6568.$$

Como a solução é positiva temos que $u_1 = 1{,}6568$. Continuando,

$$y(1) \approx u_2 = u_1 + \frac{h}{2}(f(t_1, u_1) + f(t_2, u_2)) = 1{,}3137 - 0{,}25u_2^2.$$

Resolvendo a equação $0{,}25u_2^2 + u_2 - 1{,}3137 = 0$, temos

$$0{,}25u_2^2 + u_2 - 1{,}3137 = 0 \Rightarrow u_2 = -5{,}0422 \text{ ou } u_2 = 1{,}0422.$$

Como a solução é positiva temos que $y(1) \approx u_2 = 1{,}0422$.

7.4 Estabilidade absoluta

A convergência dos métodos numéricos é verificada quando h puder ser escolhido arbitrariamente pequeno. No entanto, quando consideramos a aplicação de um método numérico consideramo-la com um h fixo. Este facto pode levar a que, especialmente se os intervalos de integração forem muito grandes, o método numérico dê uma solução que em nada corresponda à solução exacta do problema.

Por exemplo, em fenómenos dissipativos, isto é, onde as soluções do problema tendam para zero quando a variável independente tende para infinito, é muito frequente verificar que certos métodos numéricos produzem soluções oscilatórias, oscilações essas que não estão presentes na solução exacta.

Considere-se, por exemplo, a aplicação do método de Euler progressivo ao problema teste

$$\begin{cases} y' &= \lambda y, \qquad t \in]0, +\infty[\\ y(0) &= 1 \end{cases}, \tag{7.10}$$

com λ um número real negativo. A solução exacta deste problema é $y(t) = e^{\lambda t}$. Como $\lambda < 0$, tem-se que $y(t)$ tende para zero quando t tende para infinito.

Consideremos agora a solução numérica dada pelo método de Euler explícito. Temos, sucessivamente, $u_0 = 1$,

$$u_{i+1} = u_i + h\lambda u_i = (1 + h\lambda)u_i = (1 + h\lambda)^{i+1}u_0 = (1 + h\lambda)^{i+1}.$$

Assim sendo, a solução numérica tende para zero com o número de iterações, isto é,

$$\lim_{i \to +\infty} u_i = 0, \tag{7.11}$$

se e só se

$$|R(h\lambda)| < 1,$$

com $R(h\lambda) = 1 + h\lambda$. Temos então que

$$|R(h\lambda)| < 1 \Leftrightarrow -1 < 1 + h\lambda < 1 \Leftrightarrow h \in \left]0, \frac{2}{|\lambda|}\right[.$$

Um método numérico diz-se absolutamente estável se, quando aplicado ao problema teste (7.10), a sua solução numérica verifica (7.11). Podemos então dizer que o método de Euler explícito é condicionalmente absolutamente estável pois é absolutamente estável se e só se $h \in \left]0, \frac{2}{|\lambda|}\right[$.

Num problema geral (diferente do problema teste) a propriedade da estabilidade absoluta corresponde à garantia do controlo das oscilações quando t cresce.

Vamos agora considerar as importantes noções de intervalo e região de estabilidade absoluta. Consideremos um método numérico, explícito ou implícito, aplicado ao problema teste (7.10). É possível mostrar que esse método se pode escrever na forma

$$u_{i+1} = R(z)u_i, \qquad \text{com} \quad z = \lambda h.$$

O intervalo de estabilidade absoluta é definido por

$$I_{EA} = \{z \in \mathbb{R} : |R(z)| < 1\}.$$

Note-se que um método numérico é incondicionalmente absolutamente estável se e só se o intervalo $]-\infty, 0[$ estiver contido no seu intervalo de estabilidade absoluta.

No caso de se considerar, no problema teste, $\lambda \in \mathbb{C}$, com $\mathrm{Re}(\lambda) < 0$, podemos definir a região de estabilidade absoluta de um método numérico como sendo

$$R_{EA} = \{z \in \mathbb{C} : |R(z)| < 1\}.$$

Os métodos numéricos que possuem regiões de estabilidade absoluta que incluam o plano $\mathbb{C}^-$ dizem-se A-estáveis. Para esses métodos não é necessário impor qualquer restrição na medida do passo por forma a os tornar absolutamente estáveis. É possível demonstrar que não existem métodos de passo único explícitos A-estáveis.

7.5 Sistemas de equações diferenciais

A teoria apresentada nas secções precedentes pode ser facilmente generalizada para sistemas de equações diferenciais ordinárias de primeira ordem. Todos os métodos numéricos apresentados podem ser adaptados ao cálculo da solução aproximada do PCI

$$\begin{cases} Y'(t) &= F(t,Y), \qquad t \in]t_0, T] \\ Y(t_0) &= Y^{(0)} \end{cases}, \tag{7.12}$$

onde

$$Y(t) = \begin{bmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_N(t) \end{bmatrix}, \qquad F(t,Y) = \begin{bmatrix} F_1(t,Y) \\ F_2(t,Y) \\ \vdots \\ F_N(t,Y) \end{bmatrix}.$$

Os métodos numéricos irão, neste caso, determinar aproximações $U^{(i)}$ para $Y(t_i)$. O método de Euler progressivo, por exemplo, é dado por

$$U^{(i+1)} = U^{(i)} + hF(t_i, U^{(i)}), \quad i = 0, \ldots, n, \qquad U^{(0)} = Y^{(0)},$$

com $h = \frac{T-t_0}{n}$ a medida do passo.

Equações diferenciais de ordem superior a um. Uma situação importante onde surgem sistemas de equações diferenciais é quando pretendemos resolver uma equação diferencial de ordem superior a um. Note-se que qualquer equação diferencial de ordem N pode ser escrita como um sistema de N equações diferenciais de primeira ordem. A forma como essa passagem se processa é bastante simples e pode ser facilmente compreendida com a ajuda de um exemplo.

Exemplo 7.1 Consideremos o problema de condição inicial $y''-3y'+2y=0$, $y(0)=y'(0)=1$. Efectuando a mudança de variável $z=y'$ obtemos o problema de condição inicial de primeira ordem

$$\left\{\begin{array}{rcl} y'(t) &=& z \\ z'(t) &=& 3z-2y \\ y(0) &=& 1 \\ z(0) &=& 1 \end{array}\right. \Rightarrow \left\{\begin{array}{rcl} \left[\begin{array}{c} y \\ z \end{array}\right]'(t) &=& \left[\begin{array}{c} z \\ 3z-2y \end{array}\right] \\ \left[\begin{array}{c} y \\ z \end{array}\right](0) &=& \left[\begin{array}{c} 1 \\ 1 \end{array}\right] \end{array}\right. .$$

Exercício 7.8 Converta num sistema de equações diferenciais de primeira ordem o problema

$$y'''-0{,}1(1-y^2)y'+y=0, \qquad y(0)=1,\ y'(0)=y''(0)=0.$$

Resolução: Efectuando a mudança de variável $z=y'$ e $w'=y''$ obtemos o problema de condição inicial de primeira ordem

$$\left\{\begin{array}{rcl} y'(t) &=& z \\ z'(t) &=& w \\ w'(t) &=& 0{,}1(1-y^2)z-y \\ y(0) &=& 1 \\ z(0) &=& 0 \\ w(0) &=& 0 \end{array}\right. .$$

Exercício 7.9 Considere a equação diferencial $y''+4ty'+2y^2=0$ com condições iniciais $y(0)=1$ e $y'(0)=0$. Com $h=0{,}1$, utilize o método de Euler progressivo para obter aproximações para $y(0{,}2)$ e $y'(0{,}2)$.

Resolução: Seja $z=y'$. Assim o nosso problema é equivalente a

$$\left\{\begin{array}{rcl} y'(t) &=& z \\ z'(t) &=& -4tz-2y^2 \\ y(0) &=& 1 \\ z(0) &=& 0 \end{array}\right. \Rightarrow \left\{\begin{array}{rcl} \left[\begin{array}{c} y \\ z \end{array}\right]'(t) &=& \left[\begin{array}{c} z \\ -4tz-2y^2 \end{array}\right] \\ \left[\begin{array}{c} y \\ z \end{array}\right](0) &=& \left[\begin{array}{c} 1 \\ 0 \end{array}\right] \end{array}\right. .$$

Seja

$$F(t,Y)=\left[\begin{array}{c} z \\ -4tz-2y^2 \end{array}\right], \qquad \text{com} \quad Y=\left[\begin{array}{c} y \\ z \end{array}\right]$$

e

$$Y^{(0)}=\left[\begin{array}{c} y \\ z \end{array}\right](0)=\left[\begin{array}{c} 1 \\ 0 \end{array}\right].$$

Considerando o método de Euler progressivo temos

$$\left[\begin{array}{c} y \\ z \end{array}\right](0)=U^{(0)} = Y^{(0)}=\left[\begin{array}{c} 1 \\ 0 \end{array}\right]$$

$$\left[\begin{array}{c} y \\ z \end{array}\right](0{,}1)\approx U^{(1)} = U^{(0)}+hF\left(t_0,U^{(0)}\right)=\left[\begin{array}{c} 1 \\ -0{,}2 \end{array}\right]$$

$$\left[\begin{array}{c} y \\ z \end{array}\right](0{,}2)\approx U^{(2)} = U^{(1)}+hF\left(t_1,U^{(1)}\right)=\left[\begin{array}{c} 0{,}98 \\ -0{,}392 \end{array}\right].$$

Temos assim que $y(0{,}2)\approx 0{,}98$ e $y'(0{,}2)\approx -0{,}392$.

Exercício 7.10 Adapte o Algoritmo 7.1 a sistemas de equações diferenciais.

7.6 Métodos de Runge-Kutta

O método mais simples para aproximar a solução do PCI (7.2) é o método (7.7), descrito por Euler, em 1768, na sua obra *Institutiones Calculi Integralis*. É um método muito simples de entender e de programar mas, como se irá ver na próxima secção, pouco preciso. Por exemplo, se pretendermos uma precisão de, digamos, 6 casas decimais, o método de Euler necessita de aproximadamente um milhão de passos.

Se usarmos outros métodos de Taylor, a precisão pode ser aumentada. A grande desvantagem destes métodos reside no facto de termos necessidade de calcular muitas derivadas da função f para obter métodos precisos. Esse cálculo, além de muito fastidioso, torna impraticável a aplicação de tais métodos na resolução de (7.2) quando a função f tem uma expressão analítica complicada.

Uma alternativa a esses métodos foi dada por Carl David Tolmé Runge (1856-1927), em 1875, e que consistia em, partindo do conhecimento de $y(t_0)$, considerar

$$y(t_0+h) \approx y_0 + hf\left(t_0+\frac{h}{2}, y\left(t_0+\frac{h}{2}\right)\right);$$

mas, que valor atribuir a $y\left(t_0+\frac{h}{2}\right)$? A sugestão de Runge foi a de considerar o método de Euler com passo $\frac{h}{2}$. A aplicação sucessiva deste processo permitiu a Runge definir o seguinte método iterativo:

$$\begin{array}{rcl} k_1 & = & f(t_i, u_i), \\ k_2 & = & f\left(t_i+\frac{h}{2}, u_i+\frac{h}{2}k_1\right), \\ u_{i+1} & = & u_i + hk_2, \end{array} \tag{7.13}$$

com $u_i \approx y(t_i)$. Como veremos este método, apesar de recorrer ao método de Euler, vai ser mais preciso e não necessita de calcular derivadas de f. A generalização desta ideia deu origem à seguinte definição, cuja autoria é partilhada com Martin Wilhelm Kutta (1867-1944).

Definição 7.4 (Métodos de Runge-Kutta) *Seja s um número inteiro e $a_{21}, a_{31}, a_{32}, \ldots, a_{s1}, \ldots, a_{s,s-1}$, $c_2, c_3, \ldots, c_s$, $b_1, b_2, \ldots, b_s$, coeficientes reais. O método*

$$\begin{array}{rcl} k_1 & = & f(t_i, u_i), \\ k_2 & = & f(t_i + c_2 h, u_i + a_{21} h k_1), \\ k_3 & = & f(t_i + c_3 h, u_i + a_{31} h k_1 + a_{32} h k_2), \\ & \vdots & \\ k_s & = & f(t_i + c_s h, u_i + a_{s1} h k_1 + a_{s2} h k_2 + \cdots + a_{s,s-1} h k_{s-1}), \\ u_{i+1} & = & u_i + h(b_1 k_1 + b_2 k_2 + \cdots + b_s k_s), \end{array}$$

é chamado método de Runge-Kutta explícito de s etapas *para o PCI (7.2).*

Usualmente considera-se

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \qquad i = 2, 3, \ldots, s. \tag{7.14}$$

Uma notação muito usada na prática para os métodos de Runge-Kutta foi apresentada por John Charles Butcher (1933-), em 1964, e é dada pelo seguinte quadro, designado por quadro de Butcher:

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} .$$

Antes de continuarmos, notemos que os métodos de Runge-Kutta constituem uma excelente ideia. A única solução do PCI bem posto (7.2) é uma curva integral em $\mathbb{R}^2$. No entanto, devido aos erros cometidos, a solução numérica vai ser afectada pelo comportamento das curvas integrais vizinhas. É assim importante conhecer o comportamento de toda a família de curvas integrais e não apenas o de uma única curva.

Os método de Runge-Kutta usam, deliberadamente, informação de várias curvas integrais em simultâneo. A título de exemplo considere-se o método de três etapas

$$\begin{aligned} k_1 &= f(t_i, u_i), \\ k_2 &= f(t_i + c_2 h, u_i + c_2 h k_1), \\ k_3 &= f(t_i + c_3 h, u_i + (c_3 - a_{32}) h k_1 + a_{32} h k_2), \\ u_{i+1} &= u_i + h(b_1 k_1 + b_2 k_2 + b_3 k_3). \end{aligned}$$

Para determinar a solução numérica do PCI (7.2) por este método, começa-se pelo ponto (t_i, u_i) e aplica-se um passo do método de Euler com passo $c_2 h$. Seguidamente, calcula-se o valor de k_2 como sendo o vector derivada no ponto obtido. Temos assim dois valores para a derivada: k_1 e k_2; iremos usar uma média pesada entre estes dois valores,

$$(c_3 - a_{3,2}) h k_1 + a_{3,2} h k_2,$$

numa nova aplicação do método de Euler, a partir do ponto (t_i, u_i), com passo $c_3 h$. Calculando a derivada novamente obtém-se o valor de k_3. O último passo do algoritmo é mais uma aplicação do método de Euler, a partir do ponto (t_i, u_i), com passo h.

Exercício 7.11 Considere o problema de condição inicial

$$\begin{cases} y'(t) = t y^2 \\ y(1) = 2 \end{cases} .$$

Determine um valor aproximado para $y(1{,}1)$, usando o método de Heun, devido a Karl Heun (1859-1929), dado por

$$\begin{aligned} k_1 &= f(t_i, u_i), \\ k_2 &= f\left(t_i + h, u_i + h k_1\right), \\ u_{i+1} &= u_i + \tfrac{h}{2}(k_1 + k_2), \end{aligned} \tag{7.15}$$

com $h = 0{,}05$.

Resolução: Seja $f(t, y) = t y^2$. Temos que

$$\begin{aligned} y(1) = u_0 &= y_0 = 2 \\ y(1{,}05) \approx u_1 &= u_0 + \tfrac{h}{2}(k_1 + k_2) = 2 + 0{,}025(k_1 + k_2). \end{aligned}$$

Por outro lado

$$\begin{aligned} k_1 &= f(t_0, u_0) = f(1, 2) = 4 \\ k_2 &= f(t_0 + h, u_0 + h k_1) = f(1{,}05, 2{,}2) = 5{,}082. \end{aligned}$$

Assim, $y(1{,}05) \approx u_1 = 2{,}22705$. Continuando a aplicação do método

$$y(1{,}1) \approx u_2 = u_1 + \frac{h}{2}(k_1 + k_2) = 2{,}22705 + 0{,}025(k_1 + k_2).$$

Para este segundo passo temos que voltar a calcular k_1 e k_2. Assim,

$$\begin{aligned} k_1 &= f(t_1, u_1) = f(1{,}05, 2{,}22705) = 5{,}207739 \\ k_2 &= f(t_1 + h, u_1 + hk_1) = f(1{,}1, 2{,}487437) = 6{,}806077. \end{aligned}$$

Logo, $y(1{,}1) \approx u_2 = 2{,}5273954$.

Um método de Runge-Kutta (de quarta ordem) muito famoso é dado por

$$k_1 = f(t_i, u_i), \qquad k_2 = f(t_i + \frac{h}{2}, u_i + \frac{h}{2}k_1),$$

$$k_3 = f(t_i + \frac{h}{2}, u_i + \frac{h}{2}k_2), \qquad k_4 = f(t_i + h, u_i + hk_3),$$

$$u_{i+1} = u_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

O seguinte algoritmo permite determinar a solução do PCI (7.2) em $t = T$, usando este método de Runge-Kutta.

Algoritmo 7.2 Método de Runge-Kutta

Dados: n, t_0, T e y_0
$h := \frac{T - t_0}{n}$
$t := t_0$
$u := y_0$
Para i de 1 até n fazer
  $k_1 := f(t, u)$
  $k_2 := f(t + 0{,}5h, u + 0{,}5hk_1)$
  $k_3 := f(t + 0{,}5h, u + 0{,}5hk_2)$
  $k_4 := f(t + h, u + hk_3)$
  $u := u + h(k_1 + 2k_2 + 2k_3 + k_4)/6$
  $t := t + h$
Resultado: $y(T) \approx u$

Exercício 7.12 Construa um algoritmo que permita determinar a solução do PCI (7.2) em $t = T$, usando um método de Runge-Kutta explícito de s etapas qualquer.

O método de Heun é dado por (7.15). Vamos determinar qual o seu erro local e, consequentemente, qual a sua ordem. Atendendo à definição de erro local temos que $hT_i = y(t_i) - u_i^*$, com u_i^* a solução numérica obtida pelo método supondo $u_{i-1} = y(t_{i-1})$. Desenvolvendo $y(t_i)$ em série de Taylor em torno do ponto t_{i-1} temos,

$$y(t_i) = y(t_{i-1}) + hf(t_{i-1}, y(t_{i-1})) + \frac{h^2}{2}\frac{df}{dt}(t_{i-1}, y(t_{i-1})) + \frac{h^3}{6}\frac{d^2f}{dt^2}(t_{i-1}, y(t_{i-1})) + \cdots.$$

Por outo lado, considerando o desenvolvimento de u_i^*, recorrendo à série de Taylor para duas variáveis, temos

$$\begin{aligned} u_i^* &= y(t_{i-1}) + \frac{h}{2}(f(t_{i-1}, y(t_{i-1})) + f(t_{i-1}, y(t_{i-1})) + h(f_t + f_y f)(t_{i-1}, y(t_{i-1})) + \\ &\qquad \frac{h^2}{2}(f_{tt} + 2f f_{ty} + f^2 f_{yy})(t_{i-1}, y(t_{i-1})) + \cdots). \end{aligned}$$

Subtraindo membro a membro, temos

$$hT_i = \frac{h^3}{12}(f_{tt} + 2f f_{ty} + f^2 f_{yy} - 2f_t f_y - 2f f_y^2)(t_{i-1}, y(t_{i-1})) + \cdots.$$

Assim sai que

$$T_i = \frac{h^2}{12}(f_{tt} + 2f f_{ty} + f^2 f_{yy} - 2f_t f_y - 2f f_y^2)(\xi, y(\xi)), \qquad \xi \in]t_{i-1}, t_i[.$$

Como $T_i = \mathcal{O}(h^2)$ concluímos que o método de Heun tem ordem 2.

Exercício 7.13 Mostre que o método de Heun (7.15), aplicado à resolução do PCI (7.2), é convergente.

Resolução: Atendendo à definição do método de Heun temos que este pode ser dado pela expressão $u_{i+1} = y_i + h\phi(t_i, u_i; h)$, com

$$\phi(t, y; h) = \frac{1}{2}(f(t, y) + f(t + h, y + hf(t, y))).$$

Para provar que o método é convergente vamos provar que é consistente e estável.

1. Consistência. Provámos que o método de Heun tem ordem dois e, assim sendo, é consistente. Poderiamos ainda provar a consistência provando que $\phi(t, y; 0) = f(t, y)$. De facto,

$$\phi(t, y; 0) = \frac{1}{2}(f(t, y) + f(t, y)) = f(t, y).$$

2. Estabilidade. Para provar que o método é estável vamos provar que $\phi(t, y; h)$ é lipschitziana, na variável y, em $D = \{(t, y; h) : a \leq t \leq b, y \in \mathbb{R}, 0 \leq h \leq h_0\}$. Seja L a constante de Lipschitz de $f(t, y)$ na variável y. Então

$$\begin{aligned} |\phi(t, y_1; h) - \phi(t, y_2; h)| &= \left| \frac{1}{2}(f(t, y_1) + f(t + h, y_1 + hf(t, y_1))) \right. \\ &\qquad \left. - \frac{1}{2}(f(t, y_2) + f(t + h, y_2 + hf(t, y_2))) \right| \\ &\leq \frac{1}{2}(L|y_1 - y_2| + L|y_1 + hf(t, y_1) - y_2 - hf(t, y_2)|) \\ &\leq L|y_1 - y_2| + \frac{1}{2}hL^2|y_1 - y_2| \\ &\leq \left(L + \frac{1}{2}hL^2\right)|y_1 - y_2|. \end{aligned}$$

Assim $\phi(t, y; h)$ satisfaz a condição de Lipschitz, na variável y, em D sendo a sua constante de Lipschitz dada por $K = \left(L + \frac{1}{2}h_0L^2\right)$. Finalmente, tanto ϕ como f são contínuas em D. Está assim provada a estabilidade do método.

Exercício 7.14 Considere a equação diferencial $y'' + 4ty' + 2y^2 = 0$ com condições iniciais $y(0) = 1$ e $y'(0) = 0$. Com $h = 0{,}1$, utilize o método de Heun para obter aproximações para $y(0{,}2)$ e $y'(0{,}2)$.

Resolução: Seja $z = y'$. Assim o nosso problema é equivalente a

$$\begin{cases} y'(t) &= z \\ z'(t) &= -4tz - 2y^2 \\ y(0) &= 1 \\ z(0) &= 0 \end{cases} \Rightarrow \begin{cases} \begin{bmatrix} y \\ z \end{bmatrix}'(t) &= \begin{bmatrix} z \\ -4tz - 2y^2 \end{bmatrix} \\ \begin{bmatrix} y \\ z \end{bmatrix}(0) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{cases}.$$

Seja

$$F(t, Y) = \begin{bmatrix} z \\ -4tz - 2y^2 \end{bmatrix}, \qquad \text{com} \quad Y = \begin{bmatrix} y \\ z \end{bmatrix} \quad \text{e} \quad Y^{(0)} = \begin{bmatrix} y \\ z \end{bmatrix}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Considerando o método de Heun temos

$$\begin{aligned} \begin{bmatrix} y \\ z \end{bmatrix}^{(0)} = U^{(0)} &= Y^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ \begin{bmatrix} y \\ z \end{bmatrix}(0{,}1) \approx U^{(1)} &= U^{(0)} + \frac{h}{2}(K_1 + K_2), \end{aligned}$$

onde

$$\begin{aligned} K_1 &= F\left(t_0, U^{(0)}\right) = \begin{bmatrix} 0 \\ -2 \end{bmatrix} \\ K_2 &= F\left(t_0 + h, U^{(0)} + hK_1\right) = \begin{bmatrix} -0{,}2 \\ -1{,}92 \end{bmatrix}. \end{aligned}$$

Logo

$$\begin{bmatrix} y \\ z \end{bmatrix}(0{,}1) \approx U^{(1)} = \begin{bmatrix} 0{,}99 \\ -0{,}196 \end{bmatrix}.$$

Continuando a aplicação do método temos

$$\begin{bmatrix} y \\ z \end{bmatrix}(0{,}2) \approx U^{(2)} = U^{(1)} + \frac{h}{2}(K_1 + K_2),$$

onde

$$\begin{aligned} K_1 &= F\left(t_1, U^{(1)}\right) = \begin{bmatrix} -0{,}196 \\ -1{,}8818 \end{bmatrix}, \\ K_2 &= F\left(t_1 + h, U^{(1)} + hK_1\right) = \begin{bmatrix} -0{,}38418 \\ -1{,}6335 \end{bmatrix}. \end{aligned}$$

Logo

$$\begin{bmatrix} y \\ z \end{bmatrix}(0{,}2) \approx U^{(2)} = \begin{bmatrix} 0{,}988059 \\ -0{,}371765 \end{bmatrix}.$$

Temos assim que $y(0{,}2) \approx 0{,}988059$ e $y'(0{,}2) \approx -0{,}371765$.

7.7 Problemas com condições de fronteira

Neste capítulo, até ao momento, estudámos as equações diferenciais ordinárias no contexto dos sistema dinâmicos em que a variável independente natural é o tempo (nem sempre assim é). Vamos agora considerar o estudo orientado para regimes estacionários em que o objectivo consiste em determinar a distribuição espacial de uma grandeza.

Exemplo 7.2 Um problema comum em engenharia civil tem a ver com a deflexão de uma barra de secção rectangular sujeita a uma carga uniforme quando os extremos estão fixos. A equação diferencial que serve de modelo a esta situação física é da forma

$$w'' = \frac{S}{EI}w + \frac{qx}{2EI}(x - L),$$

onde $w = w(x)$ é a deflexão no ponto que dista x do extremo esquerdo da barra, e L, q, E, S e I representam, respectivamente, o comprimento da barra, a intensidede da carga uniforme, o módulo da elasticidade, a tensão nos extremos e o momento central de inércia. Uma vez que os extremos da barra estão fixos, temos associadas a esta equação diferencial as equações de fronteira

$$w(0) = w(L) = 0.$$

Quando a barra é feita de material uniforme EI é uma constante e como tal a solução da equação é imediata. Caso contrário $I = I(x)$ e temos que usar métodos numéricos para determinar uma aproximação para a solução.

Os problemas físicos que dependem de uma posição no espaço em vez de um instante no tempo são muitas vezes descritos em termos de equações diferenciais com condições impostas em mais do que um ponto: problemas com condições de fronteira (PCF). Os PCF que iremos considerar nesta secção envolvem uma equação diferencial ordinária de segunda ordem

$$y'' = f(x, y, y'), \qquad x \in]a, b[, \tag{7.16}$$

e as condições de fronteira

$$\begin{cases} \alpha_1 y(a) + \beta_1 y'(a) = \gamma_1 \\ \alpha_2 y(b) + \beta_2 y'(b) = \gamma_2 \end{cases}, \tag{7.17}$$

com $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}$, $i = 1, 2$. Estas condições de fronteira podem ser de três tipos:

1. Dirichelet, em homenagem a Johann Peter Gustav Lejeune Dirichlet (1805-1859), se $\beta_1 = \beta_2 = 0$;
2. Neumann, em homenagem a John von Neumann (1903-1957), se $\alpha_1 = \alpha_2 = 0$;
3. Robin, em homenagem a Victor Gustave Robin (1855-1897), ou mistas, se $|\alpha_1| + |\alpha_2| \neq 0$ e $|\beta_1| + |\beta_2| \neq 0$.

Quando $\gamma_1 = \gamma_2 = 0$ dizemos que as condições de fronteira são homogéneas. No caso em que a equação (7.16) é da forma

$$y'' = p(x)y' + q(x)y + r(x), \qquad x \in]a, b[, \tag{7.18}$$

dizemos que (7.16)–(7.17) é um problema com condições de fronteira linear.

Tal como no caso dos problemas de condição inicial também aqui se torna importante saber em que condições (7.16)–(7.17) tem solução única. Esse estudo foge ao âmbito deste

curso e como tal não irá ser apresentado. No entanto, para problemas com condição de fronteira lineares a teoria é mais simples e, a título ilustrativo, iremos considerar apenas o seguinte teorema, que apresentamos sem demonstração.

Teorema 7.5 *Sejam $q, r \in C([a,b])$ e $q \geq 0$. Então o PCF linear*

$$\begin{cases} -y'' + q(x)y = r(x), \quad x \in]a,b[\\ y(a) = y(b) = 0 \end{cases} \tag{7.19}$$

tem uma única solução $y \in C^2([a,b])$.

7.8 Método das diferenças finitas

Um método muito usado para determinar soluções aproximadas do PCF (7.16)–(7.17) consiste em substituir as derivadas que nela intervêm por fórmulas de diferenças finitas.

Suponhamos que o problema (7.16)–(7.17) admite uma e uma só solução e consideremos a partição

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b \tag{7.20}$$

do intervalo $[a,b]$. O método das diferenças finitas permite-nos obter aproximações u_i, $i = 0, \ldots, n$, para os valores da solução nos pontos da partição, isto é, $u_i \approx y(x_i)$, $i = 0, \ldots, n$. Por uma questão de simplificação da abordagem vamos considerar a partição (7.20) uniforme, ou seja, tal que $x_i - x_{i-1} = h$, $i = 1, \ldots, n$.

7.8.1 Caso linear

Vamos considerar o PCF linear (7.18) com condições de fronteira (7.17) de Dirichlet ($\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 0$). Este problema pode ser escrito, para cada ponto da partição (7.20), na forma

$$\begin{cases} y''(x_i) = p(x_i)y'(x_i) + q(x_i)y(x_i) + r(x_i), \quad i = 1, \ldots, n-1 \\ y(x_0) = \gamma_1, \quad y(x_n) = \gamma_2 \end{cases},$$

com $x_i = ih$, $i = 0, \ldots, n$. Substituindo as derivadas pelas fórmulas de diferenças centradas de segunda ordem

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1})}{2h} - \frac{h^2}{6}y'''(\xi_i), \qquad \xi_i \in]x_{i-1}, x_{i+1}[,$$

e

$$y''(x_i) = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} - \frac{h^2}{12}y^{(4)}(\eta_i), \qquad \eta_i \in]x_{i-1}, x_{i+1}[,$$

obtemos

$$\begin{cases} \dfrac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = p(x_i)\dfrac{y(x_{i+1}) - y(x_{i-1})}{2h} + q(x_i)y(x_i) + r(x_i) \\ \qquad\qquad -\dfrac{h^2}{12}\left[2p(x_i)y'''(\xi_i) - y^{(4)}(\eta_i)\right], \quad i = 1, \ldots, n-1 \\ y(x_0) = \gamma_1, \quad y(x_n) = \gamma_2 \end{cases}.$$

Se tomarmos $u_i \approx y(x_i)$, $i = 1, \ldots, n-1$, um método de diferenças finitas com erro $\mathcal{O}(h^2)$ pode ser definido pelo sistema linear

$$\begin{cases} \dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - p(x_i)\dfrac{u_{i+1} - u_{i-1}}{2h} - q(x_i)u_i = r(x_i), \quad i = 1, \ldots, n-1 \\ \\ u_0 = \gamma_1, \quad u_n = \gamma_2 \end{cases},$$

ou, de forma equivalente,

$$\begin{cases} \left(1 + \dfrac{h}{2}p(x_i)\right) u_{i-1} - \left(2 + h^2 q(x_i)\right) u_i + \left(1 - \dfrac{h}{2}p(x_i)\right) u_{i+1} = h^2 r(x_i), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, n-1 \\ \\ u_0 = \gamma_1, \quad u_n = \gamma_2 \end{cases}.$$

Notemos que o sistema linear obtido é da forma

$$Au = b, \tag{7.21}$$

em que $u = [u_1, u_2, \ldots, u_{n-2}, u_{n-1}]^T$,

$$A = \begin{bmatrix} -2 - h^2 q(x_1) & 1 - \frac{h}{2}p(x_1) & & & & \\ 1 + \frac{h}{2}p(x_2) & -2 - h^2 q(x_2) & 1 - \frac{h}{2}p(x_2) & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 + \frac{h}{2}p(x_{n-2}) & -2 - h^2 q(x_{n-2}) & 1 - \frac{h}{2}p(x_{n-2}) & \\ & & & 1 + \frac{h}{2}p(x_{n-1}) & -2 - h^2 q(x_{n-1}) & \end{bmatrix}$$

e

$$b = \begin{bmatrix} h^2 r(x_1) - \left(1 + \frac{h}{2}p(x_1)\right)\gamma_1 \\ h^2 r(x_2) \\ \vdots \\ h^2 r(x_{n-2}) \\ h^2 r(x_{n-1}) - \left(1 - \frac{h}{2}p(x_{n-1})\right)\gamma_2 \end{bmatrix}.$$

A questão que naturalmente se coloca é a de saber se o sistema (7.21) tem solução única. Para responder a essa questão considerere-se o seguinte exercício cuja resolução está fora do âmbito deste curso.

Exercício 7.15 Suponhamos que $A = (a_{ij})_{i,j=1}^n$ é uma matriz de ordem n, tridiagonal, com $a_{i,i-1}, a_{i,i+1} \neq 0$, para cada $i = 2, \ldots, n-1$. Se

$$|a_{11}| > |a_{12}|, \quad |a_{nn}| > |a_{n,n-1}|$$

e

$$|a_{ii}| \leq |a_{i,i-1}| + |a_{i,i+1}|, \qquad i = 2, \ldots, n-1,$$

então A é não singular.

O resultado estabelecido neste exercício permite concluir, de forma imediata, o seguinte teorema.

Teorema 7.6 *Considere-se o PCF linear (7.18) com condições de fronteira (7.17) de Dirichlet ($\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 0$) e com p, q, r funções contínuas em $[a, b]$. Se $q(x) \geq 0$, para todo o $x \in [a, b]$, então o sistema tridiagonal (7.21) tem solução única desde que $h < 2/L$, onde*

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

Exercício 7.16 Obtenha a solução aproximada do problema

$$\begin{cases} -y'' + y = x, \quad x \in]0, 1[\\ y(0) = y(1) = 0 \end{cases},$$

usando o método das diferenças finitas com uma malha uniforme de espaçamento $h = \frac{1}{n}$. Concretize para o caso $n = 4$.

Resolução: O PCF dado pode ser escrito, para cada ponto da partição (7.20) na forma

$$\begin{cases} -y''(x_i) + y(x_i) = x_i, \quad i = 1, \ldots, n-1 \\ y(x_0) = y(x_n) = 0 \end{cases},$$

com $x_i = ih$, $i = 0, \ldots, n$. Se aproximarmos y'' pela fórmula de diferenças centradas de segunda ordem (três pontos) temos

$$y''(x_i) \approx \frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}),$$

com $u_i \approx y(x_i)$, $i = 1, \ldots, n-1$. Substituindo na equação temos, em cada ponto da partição, o problema (linear) aproximado

$$\begin{cases} -(u_{i-1} - 2u_i + u_{i+1}) + h^2 u_i = ih^3, \quad i = 1, \ldots, n-1 \\ u_0 = u_n = 0 \end{cases},$$

Notemos que o sistema linear obtido é da forma $Au = b$, em que $b = (ih^3)_{i=1}^{n-1}$ e $A = (a_{ij})_{i,j=1}^{n-1}$ com

$$a_{ij} = \begin{cases} h^2 + 2, & i = j \\ -1, & j = i-1, j = i+1 \\ 0, & |j - i| > 1 \end{cases}.$$

A matriz do sistema é tridiagonal, simétrica e estritamente diagonal dominante por linhas; logo é invertível. Fica deste modo garantida a existência e unicidade de solução.

Considerando $n = 4$, ou seja $h = \frac{1}{4}$, obtemos

$$\begin{bmatrix} 2{,}03125 & -1 & 0 \\ -1 & 2{,}03125 & -1 \\ 0 & -1 & 2{,}03125 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1/128 \\ 1/64 \\ 3/128 \end{bmatrix} \Rightarrow \begin{bmatrix} 0{,}03484 \\ 0{,}05633 \\ 0{,}05004 \end{bmatrix}.$$

Pode dar-se o caso (muito frequente) das condições de fronteira não serem de Dirichlet mas de Neumann ou mistas. Suponhamos que temos o PCF

$$\begin{cases} y'' = p(x)y' + q(x)y + r(x), \quad x \in]a, b[\\ y'(a) = \gamma_1, \quad y'(b) = \gamma_2 \end{cases}. \tag{7.22}$$

Considerando, tal como para o caso anterior, a substituição das derivadas que aparecem na equação diferencial pelas fórmulas de diferenças centradas de segunda ordem obtemos

$$\left(1+\frac{h}{2}p(x_i)\right)u_{i-1}+\left(2+h^2q(x_i)\right)u_i-\left(1-\frac{h}{2}p(x_i)\right)u_{i+1}=h^2r(x_i),\quad i=1,\ldots,n-1,$$

com $u_i \approx y(x_i)$, $i=1,\ldots,n-1$. Quanto às equações de fronteira, o mais comum é considerarem-se diferenças progressivas na discretização de $y'(a)$ e regressivas na discretização de $y'(b)$. Se usarmos diferenças progressivas e regressivas com dois pontos (ordem um) obtemos

$$u_0=u_1-h\gamma_1,\qquad u_n=u_{n-1}+h\gamma_2,$$

onde $u_0 \approx y(x_0)$ e $u_n \approx y(x_n)$. Deste modo, o sistema linear a resolver difere, em relação ao caso em que considerámos condições de Dirichlet, apenas nas primeira e última linhas. Neste caso, a primeira e a última linha do sistema linear a resolver são, respectivamente,

$$\left(-1+\frac{h}{2}p(x_1)-h^2q(x_1)\right)u_1+\left(1-\frac{h}{2}p(x_1)\right)u_2=h^2r(x_1)+h\left(1+\frac{h}{2}p(x_1)\right)\gamma_1$$

e

$$\left(1+\frac{h}{2}p(x_{n-1})\right)u_{n-2}+\left(-1-\frac{h}{2}p(x_{n-1})-h^2q(x_{n-1})\right)u_{n-1}$$
$$=h^2r(x_{n-1})-h\left(1-p(x_{n-1})\frac{h}{2}\right)\gamma_2.$$

7.8.2 Caso não linear

Finalmente, façamos uma pequena abordagem ao caso não linear. Consideremos o problema não linear geral (7.16) com condições de fronteira (7.17) de Dirichlet ($\alpha_1=\alpha_2=1$, $\beta_1=\beta_2=0$). Tal como no caso linear, vamos substituir as derivadas que aparecem na equção diferencial pelas fórmulas de diferenças centradas de segunda ordem. Obtemos assim

$$\begin{cases} \dfrac{y(x_{i+1})-2y(x_i)+y(x_{i-1})}{h^2} &= f\left(x_i,y(x_i),\dfrac{y(x_{i+1})-y(x_{i-1})}{2h}-\dfrac{h^2}{6}y'''(\xi_i)\right) \\ & \quad +\dfrac{h^2}{12}y^{(4)}(\eta_i),\quad i=1,\ldots,n-1 \\ y(x_0)=\gamma_1,\quad y(x_n)=\gamma_2 \end{cases},$$

com $\xi_i,\eta_i \in]x_{i-1},x_{i+1}[$. O método de diferenças finitas que resulta quando se desprezam os termos $\mathcal{O}(h^2)$ das fórmulas de diferenças centradas e se usam as condições de fronteira é

$$\begin{cases} \dfrac{u_{i+1}-2y_i+u_{i-1}}{h^2} &= f\left(x_i,u_i,\dfrac{u_{i+1}-u_{i-1}}{2h}\right),\quad i=1,\ldots,n-1 \\ u_0=\gamma_1,\quad u_n=\gamma_2 \end{cases},$$

com $u_i \approx y(x_i)$, $i=1,\ldots,n-1$. Temos então necessidade de resolver um sistema não linear da forma

$$F(x,u)=0,$$

onde $x = [x_1, x_2, \ldots, x_{n-2}, x_{n-1}]^T$, $u = [u_1, u_2, \ldots, u_{n-2}, u_{n-1}]^T$ e

$$\begin{cases} f_1(x,u) &= -2u_1 + u_2 - h^2 f\left(x_1, u_1, \dfrac{u_2 - \gamma_1}{2h}\right) + \gamma_1 \\ f_i(x,u) &= u_{i-1} - 2u_i + u_{i+1} - h^2 f\left(x_i, u_i, \dfrac{u_{i+1} - u_{i-1}}{2h}\right), \quad i = 2, \ldots, n-2 \\ f_{n-1}(x,u) &= u_{n-2} - 2u_{n-1} - h^2 f\left(x_{n-1}, u_{n-1}, \dfrac{\gamma_2 - u_{n-2}}{2h}\right) + \gamma_2 \end{cases}.$$

Prova-se que este sistema não linear tem solução única se $h < 2L$ onde

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

A sua solução pode ser obtida, de forma aproximada, pelo método de Newton.

7.9 Problemas

7.9.1 Exercícios para resolver nas aulas

Exercício 7.17 Mostre que o problema de condição inicial

$$\begin{cases} y' &= ty \\ y(0) &= 1 \end{cases},$$

para $t \in [0, T]$, tem solução única.

Exercício 7.18 Considere o problema de condição inicial $\begin{cases} y' &= y \\ y(0) &= 1 \end{cases}$. Determine, usando o método de Euler progressivo, o valor aproximado de $y(1)$, fazendo $h = 1$, $h = 0{,}5$ e $h = 0{,}25$. Compare os resultados obtidos sabendo que a solução exacta é $y(t) = e^t$.

Exercício 7.19 (Matlab) Num circuito de voltagem aplicada E, resistência R, inductância L e capacitância C em paralelo, a corrente I satisfaz a equação diferencial

$$I' = CE'' + \frac{E'}{R} + \frac{E}{L}.$$

Suponha que $C = 0{,}3\ farad$, $R = 1{,}4\ ohm$, $L = 1{,}7\ henry$ e a voltagem é dada pela equação $E(t) = e^{-0{,}06\pi t} \sin(2t - \pi)$. Se $I(0) = 0$, determine o valor da corrente I para $t = 0{,}2j$, para $j = 1, \ldots, 5$, usando o método de Euler progressivo.

Exercício 7.20 Considere o problema de condição inicial $\begin{cases} y' &= -50y \\ y(0) &= 1 \end{cases}$ e os métodos de Euler progressivo e Euler regressivo. Usando cada um dos métodos determine a solução do problema em $t = 1$ com $h < 1$, comparando os resultados obtidos.

Exercício 7.21 Prove que os métodos de Euler progressivo e regressivo são consistentes e determine a sua ordem e erro de truncatura local.

Exercício 7.22 Determine os intervalos de estabilidade absoluta para os métodos de Euler (explícito e implícito) e para o método dos trapézios.

Exercício 7.23 Determine as regiões de estabilidade absoluta do métodos de Euler, Euler regressivo e trapézios. Conclua que os métodos dos trapézios e Euler regressivo são A-estáveis.

Exercício 7.24 (Matlab) Aplique os métodos de Euler regressivo e progressivo à resolução do problema de Cauchy

$$y' = \sin t + y, \quad t \in]0,1], \quad y(0) = 0.$$

Compare os resultados obtidos com a solução exacta $y(t) = -\frac{1}{2}(\sin(t) + \cos(t)) + \frac{1}{2}e^t$.

Exercício 7.25 (Matlab) Considere o problema de Cauchy

$$y' = -te^{-y}, \quad t \in]0,1], \quad y(0) = 0.$$

1. Aplique os métodos de Euler progressivo e regressivo com $h = 1/2, 1/2^2, ..., 1/2^{10}$.
2. Compare os resultados obtidos na alínea anterior com a solução exacta

$$y(t) = \ln(1 - \frac{t^2}{2}).$$

Exercício 7.26 (Matlab) Considere o problema de Cauchy

$$y' = -10y, \quad 0 < t \leq 2, \quad y(0) = 1,$$

cuja solução é $y(t) = e^{-10t}$. O que é que se passa quando se aplica um método de Euler com $h = 0{,}1$?

Exercício 7.27 (Matlab) Compare as soluções numéricas dos seguintes problemas com condição inicial:

1. $y' = 1 - y, \quad 0 < t \leq 2, \quad y(0) = 0, \quad$ e $\quad y' = 1 - y + 0{,}1, \quad 0 < t \leq 2, \quad y(0) = 0{,}1$
2. $y' = y, \quad 2 < t \leq 4, \quad y(2) = 0, \quad$ e $\quad y' = y + 0{,}01, \quad 2 < t \leq 4, \quad y(2) = 0{,}1.$

Exercício 7.28 (Matlab) Consideremos um corpo pontual de massa m e temperatura interna T inserido num meio ambiente de temperatura constante $T_a = 295$ K. A transferência de calor entre o corpo e o exterior pode ser descrita pela lei de *Stefan-Boltzmann*

$$v(t) = \sigma\gamma S(T^4(t) - T_a^4),$$

com t variável temporal, σ a constante de Stefan-Boltzmann ($5{,}67 \times 10^{-8}$ $\mathrm{Jm^{-2}K^{-4}s^{-1}}$), γ constante de emissividade do corpo, S a área da sua superfície e v a velocidade de transferência de calor. A taxa de variação de energia $E(t) = mCT(t)$ (onde C designa o calor específico do material que constitui o corpo) é igual, em valor absoluto, à velocidade v. Por conseguinte, fazendo $T(0) = T_0$, o cálculo de $T(t)$ exige a resolução da equação diferencial ordinária

$$\frac{dT}{dt} = -\frac{v(t)}{mC}.$$

Suponha que o corpo em questão é um cubo de lado 1 m e massa 1 Kg, $T_0 = 275$ K, $\gamma = 0{,}5$ e $C = 100$. Recorra a um método de Euler para comparar os resultados obtidos com $h = 20, 10, 5, 1$, para t a variar entre 0 e 200 segundos.

Exercício 7.29 (Matlab) Considere o seguinte problema de Cauchy

$$y' = \lambda y, \quad t > 0, \quad y(0) = 1,$$

onde λ é um número real negativo. A solução exacta é $y(t) = e^{\lambda t}$ que tende para zero quando t tende para infinito. Faça $\lambda = -1$.

1. Represente graficamente, no intervalo [0,30], as soluções obtidas para três valores diferentes de h: $h = 30/14$, $h = 30/16$ e $h = 1/2$, usando os métodos de Euler implícito e explícito.

2. Resolva a alínea anterior com o método de Crank-Nicolson, para os valores de h referidos anteriormente.

Exercício 7.30 Considere o problema de condição inicial $\left\{ \begin{array}{rcc} y' & = & ty^2 + y \\ y(1) & = & 2 \end{array} \right.$. Determine um valor aproximado para $y(1{,}1)$, usando o método de Heun.

Exercício 7.31 Considere o problema de condição inicial

$$\left\{ \begin{array}{rcl} y' & = & y - \dfrac{2t}{y} \\ y(0) & = & 1 \end{array} \right. .$$

Determine um valor aproximado para $y(0{,}8)$, usando o método de Runge-Kutta de ordem quatro:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} .$$

Exercício 7.32 Mostre que, quando o segundo membro f não depende de y, o método de Runge-Kutta de quarta ordem se reduz à aplicação da regra de Simpson.

Exercício 7.33 Mostre que o método de Runge (7.13) tem ordem dois.

Exercício 7.34 Mostre que o método de Heun é absolutamente estável se $-2 < h\lambda < 0$, em que λ é um real negativo.

Exercício 7.35 O método de Euler modificado é definido por:

$$u^*_{i+1} = u_i + hf(t_i, u_i), \qquad u_{i+1} = u_i + hf(t_{i+1}, u^*_{i+1}).$$

Determinar a condição sobre h para que este método seja absolutamente estável.

Exercício 7.36 A taxa de arrefecimento de um corpo pode ser expressa por $\frac{dT}{dt} = -k(T - T_a)$, onde T e T_a são as temperaturas do corpo e do meio circundante, respectivamente, (em graus Celsius), e k é uma constante de proporcionalidade (por minuto). Considerando que uma esfera de metal aquecida a 90^oC é mergulhada em água mantida à temperatura constante de $T_a = 20^o$C, use um método numérico para calcular quanto tempo leva a esfera a arrefecer até aos 30^oC se $k = 0{,}1$ min^{-1}.

Exercício 7.37 (Matlab) Aproxime a solução do problema

$$y'(t) = \arctan(3y) - 3y + t, \quad t > 0, \quad y(0) = 1,$$

usando o método:

1. de Euler progressivo, com $h = 2/3$ e $h = 2/3 + 0{,}1$;
2. de Euler regressivo, para os valores do passo de discretização dados na alínea anterior;
3. de Crank-Nicolson;
4. `ode23`;
5. `ode45`.

Comente os resultados obtidos.

Exercício 7.38 (Matlab) A função $y(t)$ indica a quantidade vendida de um determinado produto ao fim de t meses após ter sido introduzido no mercado. Suponha que $y(t)$ satisfaz a equação diferencial

$$\frac{dy}{dt} = \frac{2y}{t(t+1)}.$$

Ao fim do primeiro mês foram vendidas 1000 unidades daquele produto. A solução do problema é $y(t) = 4000t^2/(1+t)^2$.

1. Aproxime a solução do problema durante o primeiro ano, usando diferentes métodos numéricos.
2. Tendo em conta a evolução da venda do produto mensalmente durante o primeiro ano, aproxime o valor das vendas após 8 meses.

Exercício 7.39 Determine a solução do sistema de equações diferenciais $Y' = AY$ no instante $t = 1$, com

$$A = \begin{bmatrix} -1 & 2 \\ 1 & -4 \end{bmatrix},$$

usando o método de Euler progressivo com $h < 1$, a partir da condição inicial $Y(0) = (1, 0)$.

Exercício 7.40 (Matlab) A equação de Van der Pol

$$y'' - \mu\left(y^2 - 1\right) y' + y = 0,$$

com $\mu > 0$, é um modelo para o fluxo de corrente num tubo de vácuo com três elementos internos. Seja $\mu = 0{,}5$ e $y\left(0\right) = 0$, $y'\left(0\right) = 1$. Aproxime y e y' no intervalo temporal $[0, 30]$ usando os métodos `ode45` e `ode23s`. Repita o exercício considerando $\mu = -1000$ e $y\left(0\right) = 2$, $y'\left(0\right) = 0$ e o intervalo temporal $[0, 3000]$.

Exercício 7.41 *Determine a solução aproximada do problema*

$$\begin{cases} -y'' - 3xy = x, & x \in]0, 1[\\ y(0) = y(1) = 0 \end{cases}$$

usando o método das diferenças finitas numa malha uniforme de espaçamento $h = \frac{1}{4}$.

Exercício 7.42 *Usar o método das diferenças finitas para aproximar o problema de valores na fronteira*

$$\begin{cases} -Ty''(x) + ky(x) = w(x), \quad x \in]0,1[\\ y(0) = y(1) = 0 \end{cases},$$

onde y representa o deslocamento vertical de uma corda de comprimento 1, submetida a uma carga transversal de intensidade w por unidade de comprimento, T é a tensão e k um coeficiente associado à elasticidade da corda. No caso em que $w(x) = 1 + \sin(4\pi x)$, $T = 1$ e $k = 0{,}1$, calcular a solução correspondente a $h = \frac{1}{4}$.

7.9.2 Exercícios de aplicação à engenharia

Exercício 7.43 Um projéctil é lançado da superfície terreste com uma velocidade V. Supondo que não há arrasto a equação do movimento é

$$\nu \frac{d\nu}{dr} = -g\frac{R^2}{r^2},$$

onde ν é a velocidade à distância r do centro da Terra que tem raio R. Considerando $g = 9{,}81$ m/seg^2, $R = 6{,}37 \times 10^6$ m e $V = 15000$ m/seg, calcule a velocidade quando $r = 2R$.

Exercício 7.44 Uma solução líquida flui de forma constante ao longo de um tubo na direcção x. Alguns dos solutos contidos na solução difundem-se através da parede do tubo reduzindo a concentração z no tubo. A concentração z é dada por

$$\frac{dz}{dx} = -z(0{,}2 + \sqrt{z})e^{-0{,}03x}.$$

Se tomarmos $z = 1{,}5$ em $x = 2$ determine o valor de z em $x = 2{,}4$.

Exercício 7.45 Uma quantidade de 10 quilogramas de material é despejada num reservatório contendo 60 quilogramas de água. A concentração da solução, c (em percentagem), vem dada em função do tempo, t (em segundos), por

$$(60 - 1{,}2112c)c' = \frac{k}{3}(200 - 14c)(100 - 4c),$$

onde k, o coeficiente de transferência de massa, é igual a $0{,}0589$. A condição inicial em $t = 0$ é $c = 0$. Determine a relação entre c e t.

Exercício 7.46 A equação química irreverssível na qual duas moléculas de dicromato de potássio ($K_2Cr_2O_7$) sólido, duas moléculas de água (H_2O) e três átomos de enxofre (S) sólido dão origem a três moléculas de dióxido de enxofre (SO_2) gasoso, quatro moléculas de hidróxido de potássio (KOH) sólido e duas moléculas óxido de crómio (Cr_2O_3) sólido pode ser representada, simbolicamente, pelo esquema

$$2K_2Cr_2O_7 + 2H_2O + 3S \longrightarrow 4KOH + 2Cr_2O_3 + 3SO_2.$$

Se existirem inicialmente n_1 moléculas de $2K_2Cr_2O_7$, n_2 moléculas de H_2O e n_3 moléculas de S a equação seguinte descreve a quantidade $x(t)$ de KOH ao fim de um tempo t (em segundos)

$$x' = k\left(n_1 - \frac{x}{2}\right)^2\left(n_2 - \frac{x}{2}\right)^2\left(n_3 - \frac{3x}{4}\right)^3,$$

onde k é a velocidade da reação (constante). Se $k = 6{,}22 \times 10^{-19}$, $n_1 = n_2 = 1000$ e $n_3 = 1500$, quantas unidades de hidróxido de potássio serão formadas ao fim de 2 segundos?

Exercício 7.47 Na teoria da proliferação de uma doença contagiosa, podem ser usadas equações diferenciais relativamente elementares para prever o número de indivíduos infectados na população em cada instante, desde que sejam efectuadas simplificações apropriadas. Esta teoria foi estudada por N.T.J. Bayley em 1957 e 1967 em dois livros, um sobre matemática aplicada à medecina ('The Mathematical Approach to Biology and Medicine', John Wiley & Sons, NY, 1967) e outro sobre a teoria matemática das epidemias ('The Mathematical Theory of Epidemics', Hafner, NY, 1957).

Em particular, consideremos que todos os indivíduos numa população fixa têm uma probabilidade igual de ser infectados e que uma vez portadores da doença permanecerão sempre nessa condição. Se $x(t)$ denotar o número de indivíduos susceptíveis de contrair a doença no instante t e $y(t)$ o número de indivíduos infectados, é razoável assumir que a razão à qual o número de infectados varia é proporcional ao produto de $x(t)$ por $y(t)$ visto que a razão depende tando do número de infectados como do número de susceptíveis presentes, para cada t. Se a população for suficientemente grande para considerarmos que $x(t)$ e $y(t)$ são variáveis contínuas, o problema pode ser expresso na forma

$$y'(t) = kx(t)y(t),$$

onde k é uma constante e $x(t)+y(t) = m$ é a população total. Esta equação pode ser reescrita por forma a depender apenas de $y(t)$. Assim

$$y'(t) = ky(t)(m - y(t)). \tag{7.23}$$

1. Assumindo que $m = 100000$, $y(0) = 1000$, $k = 2 \times 10^{-6}$ e o tempo medido em dias, determine o número de indivíduos infectados ao fim de 30 dias.

2. A equação (7.23) é conhecida por equação de Bernoulli e pode ser transformada numa equação diferencial linear em $z(t)$ se efectuarmos a mudança de variável $z(t) = (y(t))^{-1}$. Usando esta técnica, determine a solução exacta $y(t)$ da equação diferencial (7.23), com as hipóteses consideradas no ponto anterior, e compare-a com a solução numérica obtida.

3. Determine $\lim_{t\to\infty} y(t)$. Este resultado está de acordo com a sua intuição?

Exercício 7.48 Consideremos um pêndulo simples constituído por uma bola uniforme de massa m e uma barra fina de comprimento l e massa negligenciável. Se considerarmos que a resistência do ar é porpocional ao quadrado da velocidade angular do pêndulo, a equação do movimeto é dada por

$$\theta'' + 2k\left(\theta'\right)^2 = -\frac{g}{l}\sin\theta,$$

sendo θ o ângulo agudo que a barra do pêndulo faz com a vertical. Considerando que em $t = 0$ se tem $\theta = \frac{\pi}{3}$ determine o valor de θ e de θ' nos instantes (em minutos) $t_i = ih$, com $h = 0{,}05$ e $i = 0, 1, \ldots, 50$.

Exercício 7.49 No exercício anterior, todos os indivíduos infectados permanecem na população ajudando a difundir a doença. Uma situação mais realista consiste em introduzir uma nova variável $z(t)$ para representar tanto o número de indivíduos que são retirados da população infectada num determinado instante t, por isolamento, como os que são tratados (e consequentemente tornados imunes) ou os que morrem. O problema posto nestes termos é, naturalmente, mais complicado mas Bayley mostrou que a solução aproximada do problema pode ser dada na forma

$$x(t) = x(0)e^{-(k_1/k_2)z(t)} \qquad \text{e} \qquad y(t) = m - x(t) - z(t),$$

onde k_1 e k_2 são, respectivamente, as taxas de crescimento de $y(t)$ e de $z(t)$, sendo $z(t)$ determinada pela equação diferencial

$$z'(t) = k_2 \left(m - z(t) - x(0)e^{-(k_1/k_2)} \right).$$

Como não é possível determinar a solução exacta deste problema, temos que recorrer à solução numérica. Assim, determine uma aproximação para $z(30)$, $y(30)$ e $x(30)$ assumindo que $m = 100000$, $x(0) = 99000$, $k_1 = 2 \times 10^{-6}$ e $k_2 = 10^{-4}$.

Exercício 7.50 O estudo de modelos matemáticos para estimar a evolução de uma população de espécies que competem entre si teve a sua origem no início do século com os trabalhos de A.J. Lotka e V. Volterra. Consideremos o problema de estimar a população constituida por duas espécies, uma das quais é predadora, cuja população no instante t é $x_2(t)$, e que se alimenta comendo a outra espécie, a que chamamos presa e cuja população é $x_1(t)$. Este problema é usualmente designado por predador-presa. Vamos assumir que a presa possui sempre uma quantidade de comida adequada e que a sua taxa de natalidade em todos os instantes é proporcional ao número de presas vivas nesse instante; isto é, a taxa de natalidade (presa) é dada por $k_1 x_1(t)$. A taxa de mortalidade das presas depende tanto do número de presas como de predadores vivos nesse instante. Por uma questão de simplicidade vamos assumir que a taxa de mortalidade (presa) é $k_2 x_1(t) x_2(t)$. A taxa de natalidade dos predadores, por outro lado, depende da quantidade de comida existente, $x_1(t)$, assim como do número de predadores existentes para fins de reprodução. Por essas razões vamos assumir que a taxa de natalidade (predador) é $k_3 x_1(t) x_2(t)$. A taxa de mortalidade dos predadores será tomada proporcionalmente ao número de predadores vivos nesse instante; isto é, a taxa de mortalidade (predador) é dada por $k_4 x_2(t)$.

A variação da população de presas e predadores pode ser dada pelas seguintes equações diferenciais

$$\begin{cases} x_1'(t) &= k_1 x_1(t) - k_2 x_1(t) x_2(t) \\ x_2'(t) &= k_3 x_1(t) x_2(t) - k_4 x_2(t) \end{cases}.$$

Assumindo que a população inicial de presas é 1000 e a de predadores 200, e que as constantes $k_1 = 3$, $k_2 = 0{,}002$, $k_3 = 0{,}0006$ e $k_4 = 0{,}5$, trace o gráfico das soluções deste problema e descreva o fenómeno físico representado. Será que o problema possui alguma solução estável? Se sim, para que valores de x_1 e x_2 é que tal acontece?

Exercício 7.51 Num livro intitulado 'Looking at History Through Mathematics', MIT Press, Cambridge MA, 1968, N. Rashevsky considerou um modelo para um problema envolvendo o evolução de não conformistas na sociedade. (Conformista é a pessoa que adopta ou segue o conformismo (anglicanismo).) . Suponhamos que uma sociedade tem uma população de $x(t)$ indivíduos no instante t, em anos, e que todos os não conformistas que acasalam com outros não conformistas têm uma descendência que também é não conformista. Por outro lado, para todas as outras descendências, existe uma proporção fixa r que são ainda não conformistas. Se as taxas de natalidade e mortalidade para todos os indivíduos se assumir como sendo as constantes n e m, respectivamente, e se conformistas e não conformistas acasalarem de forma aleatória, o problema pode ser expresso pelas equações diferenciais

$$\left\{ \begin{array}{rcl} x'(t) & = & (n-m)x(t) \\ y'(t) & = & (n-m)y(t) + rn(x(t) - y(t)) \end{array} \right. ,$$

onde $y(t)$ denota o número de não conformistas na população no instante t.

1. Se a variável $p(t) = y(t)/x(t)$ for introduzida para representar a proporção de não conformistas na sociedade no instante t, mostre que o sistema de equações diferenciais se reduz a
$$p'(t) = rn(1 - p(t)).$$

2. Assumindo que $p(0) = 0{,}01$, $n = 0{,}002$, $m = 0{,}015$ e $r = 0{,}1$, aproxime a solução $p(t)$ para os primeiros 50 anos.

3. Resolva a equação diferencial para $p(t)$ de forma exacta, e compare o resultado com a solução numérica.

Bibliografia

[1] Richard L. Burden e J. Douglas Faires, *Numerical Analysis*, Cengage Learning, 2011.

[2] Rainer Kress, *Numerical Analysis*, Spinger, 1998.

[3] Cleve Moler, *Numerical Computing with Matlab*, SIAM, 2004.

[4] Heitor Pina, *Métodos Numéricos*, McGraw Hill, Lisboa, 1995.

[5] Alfio Quarteroni e Fausto Saleri, *Cálculo Científico com o Matlab e o Octave*, Spinger, 2007.

Apêndice A

Projectos

O projecto consiste num pequeno relatório de não mais de cinco páginas sobre um assunto. Do relatório não devem fazer parte listagens de programas nem *outputs* directos das execuções dos programas.

Os critérios de avaliação serão os seguintes:

- Descrição do problema (clara e sucinta).
- Identificação dos métodos numéricos envolvidos na resolução (o que pode incluir alguma explicação se o método não foi explicado nas aulas).
- Implementação desses métodos em MATLAB.
- Execução dos programas em MATLAB em exemplos práticos (poucos mas relevantes).
- Análise dos resultados numéricos obtidos.

O trabalho realizado deve ser submetido por correio electrónico na forma de um ficheiro *zipado* com a designação `projectoX.zip` onde `X` deve ser substituído pelo dígito do número do projecto. O ficheiro zip deve incluir: todos os ficheiros MATLAB usados, um ficheiro pdf com o relatório, denominado `relatorioX.pdf`, e um ficheiro ascii, denominado `README`, contendo uma descrição sumária de todos os ficheiros enviados.

Projecto 1

Dado o sistema linear $Ax = b$ onde $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n(\mathbb{R})$ e $b \in \mathbb{R}^n$, uma classe de métodos iterativos para aproximar a solução exacta do problema pode escrever-se na forma

$$x^{(k+1)} = Bx^{(k)} + g, \quad x^{(0)} \text{ dado.}$$

Definindo $A = D - L - U$, onde

$$D = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix},$$

$$L = \begin{bmatrix} 0 & & & \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_{n-1,n} \\ & & & 0 \end{bmatrix}$$

os métodos de Jacobi e Gauss-Seidel pertencem a esta classe de métodos fazendo, respectivamente,

$$B_J = D^{-1}(L+U) \quad \text{e} \quad g_J = D^{-1}b$$

e

$$B_{GS} = (D-L)^{-1}U \quad \text{e} \quad g_{GS} = (D-L)^{-1}b$$

Em certos problemas, estes métodos são divergentes ou têm convergência demasiado lenta. Uma estratégia para contornar estas dificuldades é a introdução de um *parâmetro de relaxação*, ω. Consideremos os métodos definidos por

- método de Jacobi com relaxação (JR):
  $B_{JR}(\omega) = \omega B_J + (1-\omega)I, \quad g_{JR}(\omega) = \omega g_J$
- método de Gauss-Seidel com relaxação (GSR):
  $B_{GSR}(\omega) = (I - \omega D^{-1}L)^{-1}[(1-\omega)I + \omega D^{-1}U], \quad g_{GSR}(\omega) = (\frac{1}{\omega}D - L)^{-1}b$
- método de Gauss-Seidel com relaxação simétrico (GSRS):
  $B_{GSRS}(\omega) = (D-\omega U)^{-1}(\omega L + (1-\omega)D)(D-\omega L)^{-1}(\omega U + (1-\omega)D), \quad g_{GSRS}(\omega) = \omega(2-\omega)(D-\omega U)^{-1}D(D-\omega L)^{-1}b$

Note-se que os métodos de Jacobi e Gauss-Seidel com relaxação $\omega = 1$ coincidem com os métodos de Jacobi e Gauss-Seidel.

Considere o sistema $Ax = b$, onde

$$A = \begin{bmatrix} 62 & 24 & 1 & 8 & 15 \\ 23 & 50 & 7 & 14 & 16 \\ 4 & 6 & 58 & 20 & 22 \\ 10 & 12 & 19 & 66 & 3 \\ 11 & 18 & 25 & 2 & 54 \end{bmatrix}, \quad b = \begin{bmatrix} 110 \\ 110 \\ 110 \\ 110 \\ 110 \end{bmatrix}$$

1. Poderá definir algum dos métodos acima referidos para calcular a solução do problema?
2. Fazendo variar o parâmetro de relaxação no intervalo $[-5, 5]$, calcule o raio espectral das matrizes de iteração dos métodos anteriores. Represente esta informação na forma de um gráfico.
3. Para cada um dos métodos, determine o parâmetro de relaxação, ω_{OPT}, que torna a convergência destes mais rápida.
4. Represente graficamente o erro absoluto associado a cada um dos métodos em função do número de iterações efectuado.
5. Para cada uma das funções obtidas na alínea anterior, ajuste uma curva do tipo $y = aC^k$ no sentido dos mínimos quadrados, onde a e C são constantes a determinar e k representa o número de iterações. O que representa a constante C?
6. Qual dos métodos escolheria para resolver o sistema?

Projecto 2

Uma matriz $A \in \mathcal{M}(\mathbb{R})$ pode ser factorizada como o produto de uma matriz ortogonal, Q, com uma matriz triangular superior, R. Esta factorização, chamada *factorização QR*, pode ser construída aplicando o processo de ortogonalização de Gram-Schmidt (estudado em Álgebra Linear) às colunas de A. Segundo este processo, podemos reescrever as colunas da matriz A como combinação linear de uma base ortonormada. Essa base usa-se para construir a matriz Q e os coeficientes da combinação linear usam-se para construir R.

1. Faça uma função em Matlab que dada uma matriz A, calcule os factores da factorização QR usando o algoritmo de Gram-Schmidt. A função deve ter a seguinte sintaxe: `function [Q,R] = factQR(A)`.

2. Recorrendo à função anterior, implemente uma nova função que calcule todos os valores próprios segundo o algoritmo descrito na Secção 6.2 dos apontamentos das aulas. Esta função deve admitir como argumentos de entrada

   - uma matriz A, da qual queremos calcular os valores próprios
   - uma tolerância *tol*, que deverá estar associada a um critério de paragem adequado
   - o número máximo de iterações *nmax* admitidas

3. Se aplicar o algoritmo anterior à matriz (sem o critério de paragem da tolerância)

$$\begin{bmatrix} -1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \\ -2 & 3 & -4 & 5 \\ 4 & 3 & 2 & -1 \end{bmatrix}$$

   o que acontece à estrutura da matriz das iterações? Qual a relação entre a estrutura dessa matriz e os valores próprios de A?

Projecto 3

Pretende-se calcular o valor aproximado do integral

$$\int_a^b f(x)dx$$

de uma função $f : [a, b] \longrightarrow \mathbb{R}$.

O método de integração de Romberg consiste em aplicar a regra dos trapézios em malhas uniformes convenientes e depois combinar esses resultados de uma forma inteligente para obter uma melhor aproximação para o valor do integral.

Se $h = b - a$ e $T(h)$ denotar o resultado obtido pela aplicação da regra dos trapézios composta com passo h, no intervalo $[a, b]$ então a aproximação gerada pelo método de Romberg obtém-se através do algoritmo

$$\begin{aligned} \mathcal{A}_{m,0} &= T\left(\frac{b-a}{2^m}\right), && m = 0, \ldots, n \\ \mathcal{A}_{m,q+1} &= \frac{4^{q+1}\mathcal{A}_{m,q} - \mathcal{A}_{m-1,q}}{4^{q+1} - 1}, && q = 0, \ldots, n-1 \\ & && m = q+1, \ldots, n. \end{aligned}$$

que pode ser representado através do diagrama

$$\begin{array}{ccccccccccc}
\mathcal{A}_{0,0} & & & & & & & & & & \\
 & \searrow & & & & & & & & & \\
\mathcal{A}_{1,0} & \rightarrow & \mathcal{A}_{1,1} & & & & & & & & \\
 & \searrow & & \searrow & & & & & & & \\
\mathcal{A}_{2,0} & \rightarrow & \mathcal{A}_{2,1} & \rightarrow & \mathcal{A}_{2,2} & & & & & & \\
 & \searrow & & \searrow & & \searrow & & & & & \\
\mathcal{A}_{3,0} & \rightarrow & \mathcal{A}_{3,1} & \rightarrow & \mathcal{A}_{3,2} & \rightarrow & \mathcal{A}_{3,3} & & & & \\
 & \searrow & & \searrow & & \searrow & & \searrow & & & \\
\vdots & \vdots & & \vdots & & \vdots & & \ddots & & & \\
 & \searrow & & \searrow & & \searrow & & & \searrow & & \\
\mathcal{A}_{n,0} & \rightarrow & \mathcal{A}_{n,1} & \rightarrow & \mathcal{A}_{n,2} & \rightarrow & \mathcal{A}_{n,3} & \ldots & \rightarrow & \mathcal{A}_{n,n} &
\end{array}$$

É possível mostrar, sob certas condições de regularidade da função f, que a sucessão anterior satisfaz

$$\mathcal{A}_{m,n} = \int_a^b f(x)dx + \mathcal{O}\left(\left(\frac{b-a}{2^m}\right)^{2(n+1)}\right). \tag{A.1}$$

1. Implemente uma função em MATLAB que, dado um intervalo $[a, b]$, uma função $f : [a, b] \longrightarrow \mathbb{R}$ e os índices m e n, implemente o algoritmo anterior para calcular $\mathcal{A}_{m,n}$.

2. Considere o integral
$$\int_0^\pi e^x \cos(x)dx = -\frac{e^\pi + 1}{2}.$$

   (a) Faça uma tabela com o erro cometido na aproximação do integral anterior para valores de n e m a variar entre 0 e 5.

   (b) Fixando $n = 0, 1, 2$ e fazendo m variar, faça um gráfico com o erro absoluto cometido nas três aproximações em função de h (considere m suficientemente grande tal que o menor valor de h seja superior a 10^{-6}). Qual a ordem de convergência dos três métodos? Os resultados que obtém são concordantes com a estimativa (A.1)?

   (c) Proceda de modo análogo à alínea anterior e compare as aproximações obtidas com $\mathcal{A}_{m,0}$, $\mathcal{A}_{m,1}$ e $\mathcal{A}_{m,2}$, com o erro cometido com a regra dos trapézios, Simpson e ponto médio. Estime as ordens de convergência dos métodos e com base nesse resultado, indique qual das seis regras utilizaria para aproximar o valor do integral.

   (d) Utilizando os comandos de MATLAB `tic` e `toc`, compare os seis métodos referidos na alínea anterior relativamente aos tempos de execução (*Sugestão*: represente, em função de h, o tempo que demora a executar cada um dos métodos).

Projecto 4

1. Resolva a equação integral
$$\int_0^1 (s^2 + t^2)^{1/2} u(t)dt = \frac{(s^2+1)^{3/2} - s^3}{3}$$

no intervalo $[0, 1]$ discretizando o integral usando a regra de Simpson composta com n pontos t_j igualmente distanciados, usando o mesmo n para obter os pontos igualmente distanciados s_i. Resolva o sistema linear $Ax = b$ resultante da discretização usando o método da eliminação de Gauss, considerando vários valores de n a variar entre 3 e 15, comparando os resultados com a única solução do sistema, $u(t) = t$. Que valor de n dá os melhores resultados? Pode explicar porquê?

2. Para cada valor de n considerado na alínea anterior, calcule o número de condição da matriz. Como se comporta o número de condição como função de n?

3. Resolva o sistema linear obtido na alínea (a) usando o método da regularização, que consiste em aproximar a solução do sistama linear pela solução do problema de minimização
$$\min_x(\|y - Ax\|_2^2 + \mu\|x\|_2^2),$$
onde o parâmetro μ corresponde ao peso relativo dado à norma do resíduo e à norma da solução. O problema de minimização pode ser visto como um problema de mínimos quadrados associado ao sistema
$$\begin{bmatrix} A \\ \sqrt{\mu}I \end{bmatrix} x \approx \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

4. Para cada valor de μ trace um gráfico onde os eixos coordenados são a norma da solução e a norma do resíduo. Qual a forma da curva obtida à medida que μ varia? Essa forma sugere a existência de um valor óptimo para μ?

Projecto 5

A figura mostra um indutor espiral plano, implementado em CMOS para uso em circuitos de RF. O indutor é caracterizado por quatro parâmetros-chave:

- n, o número de voltas (que é um múltiplo de 1/4, mas que não precisamos nos preocupar com isso);
- w, a largura do arame;
- d, o diâmetro interno;
- D, o diâmetro exterior.

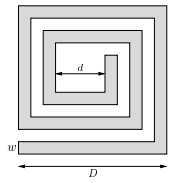

Figura A.1: Indutor espiral plano.

A indutância L de um indutor espiral plano é uma função complexa dos parâmetros n, w, d, e D. Ela pode ser encontrada através da resolução de equações de Maxwell, para

o que é necessário um considerável tempo computacional, ou fabricando o indutor e medindo a indutância. Neste problema pretende-se desenvolver um modelo simples indutância aproximada da forma

$$L = \alpha n^{\beta_1} w^{\beta_2} d^{\beta_3} D^{\beta_4},$$

onde α, β_1, β_2, β_3 e β_4 são constantes reis que caracterizam o modelo aproximado. (Como L é positivo, temos que $\alpha > 0$, mas as constantes $\beta_2, ..., \beta_4$ podem ser negativas.) Este modelo aproximado simples, se for suficientemente preciso, pode ser usado para o projecto de indutores espirais planares. O ficheiro `DadosProjecto5.m`, disponível na página das disciplina, contém dados de 50 indutores, obtidos a partir de medições.

1. Faça o download do ficheiro execute-o em Matlab na forma `[n, w, d, D, L] =DadosProjecto5`. Isso gera cinco vectores n, w, d, D, L de comprimento 50. Os i-ésimos elementos destes vectores são os parâmetros n_i, w_i (em μm), d_i (em μm), D_i (em μm) e a indutância L_i (em nH) para o indutor i. Assim, por exemplo, w_{13} dá a largura do fio do indutor 13.

2. Determine α, β_1, β_2, β_3 e β_4 tais que

   $$\hat{L}_i = \alpha n_i^{\beta_1} w_i^{\beta_2} d_i^{\beta_3} D_i^{\beta_4} \approx L_i, \qquad i = 1, ..., 50.$$

   A solução deve incluir uma descrição clara de como foram determinados os parâmetros, bem como seus valores numéricos reais.

   Note-se que não foi especificado o critério a usar para obter o modelo aproximado (isto é, o ajuste entre $\hat{L}_i$ e L_i); esse critério, no entanto, terá que ser especificado.

3. Pode definir-se o erro (em percentagem) entre $\hat{L}_i$ e L_i como

   $$e_i = 100 \frac{|\hat{L}_i - L_i|}{L_i}.$$

   Determine o erro médio para os 50 indutores, ou seja, $(e_1 + \cdots + e_{50})/50$, para o modelo considerado.

Projecto 6

A figura mostra um sinal de comprimento de 1000, corrompido com ruído. O objectivo consiste em estimar o sinal original. Este processo é chamado de reconstrução do sinal, ou de eliminação do ruído, ou alisamento. Neste problema pretende-se aplicar um método baseado no algoritmo dos mínimos quadrados.

Vamos representar o sinal corrompido como um vector x_{cor} de tamanho 1000. Os valores podem ser obtidos fazendo `xcor = DadosProjecto6`, onde `DadosProjecto6.m` é um ficheiro que pode ser obtido a partir da página da disciplina. O sinal estimado (ou seja, a variável do problema) será representado por um vector x de tamanho 1000.

A ideia do método é a seguinte. Suponhamos que o ruído no sinal é caracterizado por ser uma variável pequena e de variação rápida. Para reconstruir o sinal, decompõe-se x_{cor} em duas partes

$$x_{cor} = \hat{x} + v,$$

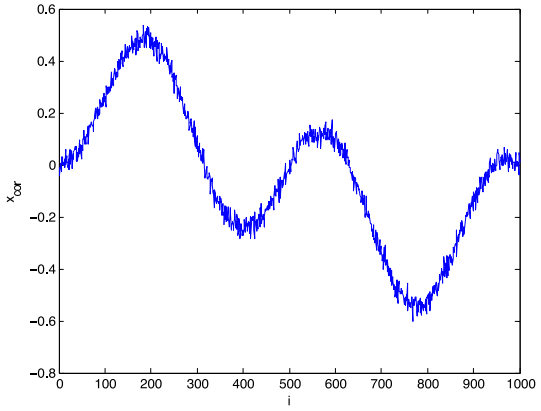

Figura A.2: Sinal com ruído.

onde v é pequena e de variação rápida e $\hat{x}$ está próximo de x_{cor} ($\hat{x} \approx x_{cor}$) e tem variação lenta ($\hat{x}_{i+1} \approx \hat{x}_i$). Pode obter-se tal decomposição escolhendo x como sendo a solução do problema dos mínimos quadrados

$$\text{minimizar } \|x - x_{cor}\|_2^2 + \mu \sum_{i=1}^{999} (x_{i+2} - x_i)^2, \tag{A.2}$$

onde μ é uma constante positiva. O primeiro termo $\|x - x_{cor}\|_2^2$ mede o quanto x se desvia de x_{cor}. O segundo termo, $\sum_{i=1}^{999}(x_{i+2} - x_i)^2$, penaliza as rápidas mudanças do sinal entre duas amostras. Ao minimizar a soma ponderada de ambos os termos, obtemos uma estimativa x que está perto de x_{cor} (ou seja, tem um pequeno valor de $\|x - x_{cor}\|_2^2$) e varia lentamente (ou seja, tem um pequeno valor de $\sum_{i=1}^{999}(x_{i+2} - x_i)^2$). O parâmetro μ é usado para ajustar o peso relativo dos dois termos. O problema (A.2) é um problema dos mínimos quadrados, pois pode ser expresso como

$$\text{minimizar } \|Ax - b\|_2^2,$$

onde

$$A = \begin{bmatrix} I \\ \sqrt{\mu} D \end{bmatrix}, \qquad b = \begin{bmatrix} x_{cor} \\ 0 \end{bmatrix},$$

e D é a matriz 999×1000 definida como

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix}.$$

A matriz é muito grande (1999×1000), mas também muito esparsa e, como tal, um bom algoritmo para o problema dos mínimos quadrados é o da factorização de Cholesky.

1. Verifique que as equações normais são dadas por

$$(I + \mu D^T D)x = x_{cor} \tag{A.3}$$

O Matlab fornece rotinas especiais para a resolução de equações lineares esparsas, e elas são utilizadas da forma seguinte. Existem dois tipos de matrizes: cheias (ou densas) e esparsas. Ao definir uma matriz, considera-se, por defeito, que ela é densa, a menos que se especifique que ela é esparsa. Pode converter-se uma matriz densa numa esparsa usando o comando `A = sparse(A)`, e um matriz esparsa numa densa usando o comando `A = full(A)`.

O comando para criar uma matriz esparsa nula de dimensão $m \times n$ é `A = sparse (m, n)`. O comando `A= speye (n)` cria a matriz identidade esparsa de dimensão $n \times n$.

2. Resolva o problema dos mínimos quadrados (A.3) com o vector x_{cor} definido em `ch9ex9.m`, para três valores de μ: $\mu = 1$, $\mu = 100$ e $\mu = 10000$. Trace os três sinais reconstruídos x.

3. Discuta o efeito de μ sobre a qualidade da estimativa x.

Projecto 7

Pretende-se simular a trajectória de uma bola de basebol do lançador (*pitcher*) para o receptor (*catcher*). Adoptando o referencial indicado na figura, as equações que descrevem o movimento da bola são

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}, \qquad \frac{d\mathbf{v}}{dt} = \mathbf{F},$$

onde $\mathbf{x}(t) = (x(t), y(t), z(t))^T$ indica a posição da bola no instante t, $\mathbf{v} = (v_x(t), v_y(t), v_z(t))^T$ a sua velocidade, e $\mathbf{F}$ o vector cujas componentes são:

$$\begin{aligned} F_x &= -F(v)vv_x + Bw(v_z \sin\phi - v_y \cos\phi), \\ F_y &= -F(v)vv_y + Bwv_x \cos\phi, \\ F_z &= -g - F(v)vv_z - Bwv_x \sin\phi, \end{aligned} \tag{A.4}$$

onde v é o módulo de $\mathbf{v}$, $B = 4{,}1 \times 10^{-4}$ uma constante normalizada, ϕ é o ângulo de lançamento, w o módulo da velocidade angular incutida à bola pelo lançador. $F(v)$ é o coeficiente de atrito, usualmente definido por

$$F(v) = 0{,}0039 + \frac{0{,}0058}{1 + e^{(v-35)/5}}.$$

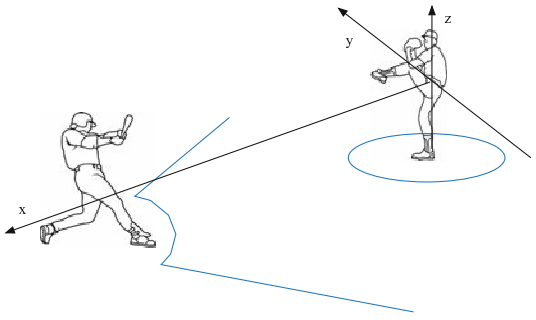

Figura A.3: Referencial para o movimento da bola de basebol.

1. Resolva o problema usando um método explícito, assumindo o valor inicial para o movimento da bola como sendo $\mathbf{v}(0) = v_0(\cos\phi, 0, \sin\phi)^T$, com $v_0 = 38$ m/s, $\phi = 1$ grau e uma velocidade angular de $180 \times 1{,}047198$ radianos por segundo.

2. Se $\mathbf{x}(0) = \mathbf{0}$, ao fim de quantos segundos (aproximadamente) a bola tocará no chão (i.e., $z = 0$)?

Projecto 8

O problema de Kepler descreve o movimento de dois corpos que se atraem mutuamente. Se escolhermos um dos corpos como o centro do sistema de coordenadas, é possível mostrar que o movimento permanece sempre no mesmo plano.

Denotando por $q = (q_1, q_2)$ a posição do segundo corpo, a Lei de Newton permite concluir, mediante uma conveniente normalização, que o movimento é dado por

$$\frac{d}{dt}\begin{bmatrix} p \\ q \end{bmatrix} = J\nabla H(p,q), \qquad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

onde a energia total do sistema (o Hamiltoniano) é

$$H(p,q) = H(p_1, p_2, q_1, q_2) = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}.$$

1. Determine a solução numérica do problema no intervalo $[0, T]$, considerando

$$p_1(0) = 0, \quad p_2(0) = \sqrt{\frac{1+e}{1-e}}, \quad q_1(0) = 1 - e, \quad q_2(0) = 0,$$

com $0 \leq e < 1$ a excentricidade (pode considerar $e = 0{,}2$, $0{,}4$ e $0{,}6$) e $T = N \times 2\pi$ (a solução tem período 2π) com $N = 10, 100, 1000$.

Considere os métodos numéricos

Euler explícito: $u_{i+1} = u_i + hF(u_i)$,

Euler simpléctico: $p_{i+1} = p_i + hf(p_i, q_{i+1}), \quad q_{i+1} = q_i + hg(p_i, q_{i+1})$,

aplicados a um sistema da forma

$$u' = \frac{d}{dt}\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} f(p,q) \\ g(p,q) \end{bmatrix} = F(u).$$

2. Compare os métodos numéricos quanto ao esforço computacional e à qualidade das órbitas obtidas.

3. O sistema solar exterior tem sido muito estudado pelos astrónomos que pretendem saber o seu comportamento para um período de tempo de aproximadamente 100 milhões de anos. Este problema é, de novo, um problema Hamiltoniano onde

$$H(p,q) = \frac{1}{2}\sum_{i=0}^{5} m_i^{-1} p_i^T p_i - G\sum_{i=1}^{5}\sum_{j=0}^{i-1} \frac{m_i m_j}{\|q_i - q_j\|}.$$

Aqui p e q são super-vectores compostos pelos vectores $p_i, q_i \in \mathbb{R}^3$. As unidades escolhidas são: massas relativas ao Sol (por forma a que a massa do Sol seja 1)

mas, para ter em conta os planetas interiores, faz-se $m_0 = 1{,}00000597682$; distâncias em UA (1 UA=149 597 879 km); tempo em dias e a constante gravitacional é $G = 2{,}95912208286 \times 10^{-4}$. Considera-se o Sol imóvel na origem e os dados para os restantes planetas são os dados na seguinte tabela referente ao dia 5 de Setembro de 1994 às 0h00.

planeta	massa	posição inicial	velocidade inicial
Júpiter	0,000954786104043	-3,5023653	0,00565429
		-3,8169847	-0,00412490
		-1,5507963	-0,00190589
Saturno	0,000285583733151	9,0755314	0,00168318
		-3,0458353	0,00483525
		-1,6483708	0,00192462
Urano	0,0000437273164546	8,3101420	0,00354178
		-16,2901086	0,00137102
		-7,2521278	0,00055029
Neptuno	0,0000517759138449	11,4707666	0,00288930
		-25,7294829	0,00114527
		-10,8169456	0,00039677
Plutão	$1/(1.3 \times 10^8)$	-15,5387357	0,00276725
		-25,2225594	-0,00170702
		-3,1902382	-0,00136504

Determine a solução do problema para um período de 200 000 dias usando métodos com passo $h = 10$ dias.

Projecto 9

Considere um ecossistema simples formado por coelhos e raposas. Os coelhos dispõem de alimentação em quantidade infinita e as raposas comem os coelhos que conseguirem apanhar. Um modelo matemático clássico para estudar a interacção de duas espécies (predador-presa), conhecido por modelo de Lotka-Volterra, consiste no seguinte sistema de duas equações diferenciais ordinárias não lineares de primeira ordem:

$$\begin{cases} x'(t) &= ax(t) - cx(t)y(t), \qquad x(0) = x_0, \\ y'(t) &= -by(t) + dx(t)y(t), \qquad y(0) = y_0, \end{cases}$$

em que $x(t)$ é o número de coelhos no instante t, $y(t)$ o número de raposas no mesmo instante e a, b, c e d são constantes positivas representando o comportamento das duas espécies. Se $c = d = 0$ as duas espécies não interaccionam, o que, nas circunstâncias deste problema, significa que os coelhos levam uma vida regalada e as raposas morrem de fome. Quando c e d forem maiores que zero, as raposas têm oportunidade de rectificar esta situação.

1. Investigue o comportamento deste ecossistema quando $a = 2$, $b = 1$ e $c = d = 0{,}01$ para vários valores iniciais das populações de coelhos e raposas, desde 2 ou 3 unidades até às centenas.

2. Verifique a existência de soluções periódicas, nomeadamente quando $x_0 = 300$, $y_0 = 150$, em que o período é aproximadamente igual a 5 unidades de tempo.

3. Investigue também se há possibilidade de extinção de alguma das espécies, ou mesmo de ambas. Considere que, se a população de uma espécie descer abaixo de 2, essa espécie se extingue (por razões óbvias!).

Projecto 10

Para calcular a forma de uma calha de escoamento, por gravidade, com o objectivo de minimizar o tempo de descarga de um determinado produto granulado (Chiarella, Charlton, Roberts (1975)), é necessário resolver as seguintes equações não lineares pelo método de Newton

$$\begin{cases} f_n(\theta_1,\ldots,\theta_N) \equiv \dfrac{\sin\theta_{n+1}}{v_{n+1}}(1-\mu w_{n+1}) - \dfrac{\sin\theta_n}{v_n}(1-\mu w_n) = 0, \quad n=1,\ldots,N-1 \\ f_N(\theta_1,\ldots,\theta_N) \equiv \Delta y \displaystyle\sum_{j=1}^{N} \tan\theta_j - X = 0 \end{cases},$$

onde

(i) $v_n^2 = v_0^2 + 2gn\Delta y - 2\mu\Delta y \displaystyle\sum_{j=1}^{n} \frac{1}{\cos\theta_j}, \qquad n=1,\ldots,N,$

e

(ii) $w_n = -\Delta y v_n \displaystyle\sum_{j=1}^{N} \frac{1}{v_j^3\cos\theta_j}, \qquad n=1,\ldots,N.$

A constante v_0 é a velocidade inicial do produto granulado, X a coordenada em x do extremo final da calha, μ a força de atrito, N o número de segmentos da calha e g a constante de gravidade. As variáveis θ_j são os ângulos que os respectivos segmentos da calha fazem com a vertical e v_j a velocidade das partículas no j-ésimo segmento da calha.

1. Resolva o sistema para $\theta = (\theta_1,\ldots,\theta_N)$ usando o método de Newton com $\mu = 0$, $X = 2$, $\Delta y = 0{,}2$, $N = 20$, $v_0 = 0$, e $g = 32$ pés/seg^2, onde os valores de v_n e w_n são dados por (i) e (ii). Aplique o método até que
$$\|\theta^{(k+1)} - \theta^{(k)}\| < 10^{-2}\text{rad}.$$

2. Conclua, numericamente, que o método tem ordem de convergência 2.

Projecto 11

Uma matriz de Hilbert de ordem n tem entradas $h_{ij} = 1/(i+j-1)$ e, como tal, tem a forma

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Para $n = 2, 3, ...$, obtenha a matriz de Hilbert de ordem n e também o vector $b = Hx$, com x o vector de dimensão n com todas as suas entradas iguais à unidade.

1. Usando o método da eliminação de Gauss (ou a factorização de Cholesky, uma vez que a matriz H é simétrica e positiva definida) resolva o sistema de equações lineares $Hx = b$, obtendo uma aproximação $\hat{x}$.

2. Calcule a norma infinito do resíduo $r = b - H\hat{x}$ e do erro $\Delta x = x - \hat{x}$, onde x é o vector de dimensão n com todas as suas entradas iguais a um. Quão grande deverá ser o valor de n por forma a que o erro seja de 100% (isto é, não existem dígitos significativos correctos na solução).

3. Determine uma estimativa para o número de condição da matriz H para cada valor de n e tente caracteriza-lo em função de n. À medida que o n varia, de que forma o número de dígitos significativos correctos da solução aproximada está relacionado com o número de condição da matriz?

Projecto 12

A intensidade de luz com comprimento de onda λ viajando através de uma grelha de difracção com n aberturas a um ângulo θ é dada por

$$I(\theta) = \left(\frac{n}{k}\right)^2 \sin^2 k,$$

onde

$$k = \frac{\pi n d \sin\theta}{\lambda}$$

e d é a distância entre cada abertura. Um laser de hélio-néon com comprimento de onda $\lambda = 632{,}8 \times 10^{-9}$ m emite uma banda estreita de luz, dada por $-10^{-6} < \theta < 10^{-6}$, através de uma grelha com 10000 aberturas separadas por 10^{-4} m.

1. Obtenha um valor aproximado para a intensidade de luz total que sai da grelha

$$\int_{-10^{-6}}^{10^{-6}} I(\theta)\, d\theta,$$

   usando os diferentes algoritmos estudados.

2. Compare os resultados obtidos, evidenciando a ordem de convergência dos métodos usados.

Projecto 13

Considere uma esfera de fio eléctrico de raio r e condutividade σ. Pretende-se calcular a densidade de distribuição da corrente $\mathbf{j}$ em função de r e t (o tempo), sabendo a distribuição inicial da densidade de carga $\rho(r)$. O problema pode ser resolvido usando a relação entre a densidade de corrente, o campo eléctrico e a densidade de carga e notando que, pela simetria do problema, $\mathbf{j}(r,t) = j(r,t)\mathbf{r}/|\mathbf{r}|$, com $j = |\mathbf{j}|$. Obtém-se assim

$$j(r,t) = \gamma(r)e^{-\sigma t/\epsilon_0}, \qquad \gamma(r) = \frac{\sigma}{\epsilon_0 r^2}\int_0^r \rho(\varepsilon)\varepsilon^2 d\varepsilon,$$

onde $\epsilon_0 = 8{,}859 \times 10^{-12}$ farad / m é a constante dieléctrica do vácuo.

1. Usando os métodos de quadratura estudados, determine a função $j(r,0)$ para $r = k/10$ m, para $k = 1, ..., 10$, $\rho(\varepsilon) = e^{\varepsilon}$ e $\sigma = 0{,}36$ W/(mK). Assegure-se que o erro de quadratura é inferior a 10^{-10}.

2. Compare os resultados obtidos, evidenciando a ordem de convergência dos métodos usados.

Projecto 14

Considere o problema de valor de fronteira

$$-\epsilon u'' + u' = 1 \text{ em }]0,1[, \qquad u(0) = u(1) = 0,$$

onde $\epsilon > 0$.

1. Verifique que a solução exacta é dada por

$$u(x) = x - \frac{e^{x/\epsilon} - 1}{e^{1/\epsilon} - 1}.$$

2. Trace o gráfico da solução para $\epsilon = 0{,}1, 0{,}01, 0{,}001$ (é necessário uma abordagem inteligente para evitar o *overflow* para o caso em que $\epsilon = 0{,}001$).

3. Calcule e trace o gráfico da solução usando um método de diferenças finitas numa rede uniforme de espaçamento $h = 1/n$.

4. Investigue a convergência do método numérico quando h decresce usando a norma máxima e a norma

$$\|u\|_p = \left(h \sum_{i=1}^{n} |u(ih)|^p \right)^{1/p},$$

quando $p = 1$ e $p = 2$. Descreva o comportamento da convergência e como varia para os três valores de ϵ considerados.

5. Seja $m = n/2$ (assumindo que n é par) e $h_1 = \bar{x}/m$, $h_2 = (1 - \bar{x})/m$, com $\bar{x} \in]0,1[$ um valor a ser definido. Considere os pontos da rede

$$x = ih_1, \quad i = 0, 1, ..., m, \qquad x_{m+i} = \bar{x} + ih_2, \quad i = 1, 2, ..., m.$$

Isto significa que se está a usar o espaçamento h_1 para os pontos

$$0 = x_0 < x_1 < \cdots < x_m = \bar{x}$$

e o espaçamento h_2 para

$$\bar{x} = x_m < x_{m+1} < \cdots < x_n = 1.$$

Para o problema em causa considere-se

$$\bar{x} = 1 - \epsilon |\log \epsilon|.$$

A malha resultante é chamada *malha de Shishkin*.

Obtenha a solução numérica para o problema dado usando o método de diferenças finitas definido nesta malha e estude a sua convergência. Certifique-se que está a definir o método numérico correctamente no ponto $x_m = \bar{x}$, onde os pontos vizinhos não estão igualmente distanciados.

Projecto 15

Uma fábrica produz efluentes que contêm um poluente cuja concentração excede os limites normais. Para reduzir o nível de poluição, a fábrica decidiu adoptar o seguinte processo de tratamento: os efluentes são primeiro enviados para um tanque contendo bactérias cuja finalidade é digerir o poluente e só depois são lançados num rio. Para estudar este processo, desenvolveu-se um modelo matemático baseado nas seguintes hipóteses:

- o tanque está equipado com um misturador que assegura que a concentração de poluente no tanque e à saída seja a mesma e, analogamente, para a concentração de bactérias;
- as bactérias reproduzem-se proporcionalmente ao alimento de que dispõem, i.e., à concentração de poluente;
- a digestão do poluente é também proporcional à concentração de bactérias.

Sejam

- V o volume do tanque (20 000 m^3);
- Q o caudal de efluentes (10 m^3/h);
- $p*(t)$ a concentração de poluente à entrada do tanque (0,01 a 0,1 g/m^3);
- $p(t)$ *idem*, no tanque à sua saída (g/m^3);
- $b(t)$ a concentração de bactérias no tanque e à sua saída (g/m^3);
- p_l o limite normal estabelecido para a concentração de poluente (5×10^{-4} g/m^3);
- R a taxa de reprodução das bactérias (1,26 m^3/gh);
- M a taxa de mortalidade das bactérias ($10^{-5}/\mathrm{h}$);
- D a taxa de digestão do poluente (0,1 m^3/gh).

As equações correspondentes ao modelo adoptado são

$$\begin{cases} V\dfrac{dp}{dt} &= Qp* - Qp - DVbp, \\ V\dfrac{db}{dt} &= RVpb - MVb - Qb. \end{cases}$$

1. Confirme que o tanque possui um volume suficiente para garantir o objectivo pretendido em regime estacionário, i.e., quando $dp/dt = 0$ e $db/dt = 0$.
2. A fábrica não opera todavia sempre nas mesmas condições pelo que o caudal de efluentes e a sua concentração em poluentes podem sofrer variações em relação aos valores em regime estacionário. Pretende-se fazer uma análise de duas situações possíveis, determinando em particular se o limite de poluição é excedido ou não.
   - $Q = 15$ m^3/h e $p* = 0{,}05$ g/m^3 durante 2 horas;
   - $Q = 10$ m^3/h e um salto instantâneo do nível de poluição de 0,01 para 0,1 g/m^3.

   Sempre que necessário usar como condições iniciais de p e b os respectivos valores em regime estacionário.

   As autoridades sanitárias estão dispostas a tolerar que a concentração de poluente exceda p_l, mas não $2p_l$, durante um período limitado de 12 horas.