# Notas de Análise Numérica

João Lopes Dias

21 de Maio de 2019

**Resumo**

Estas notas destinam-se à cadeira Análise Numérica[1] da Licenciatura em Matemática Aplicada à Economia e Gestão do ISEG - Universidade de Lisboa. Para as seguir pressupõe-se conhecimentos das áreas de álgebra linear, cálculo diferencial e integral em $\mathbb{R}^d$ e equações diferenciais ordinárias. Para a resolução de exercícios computacionais assume-se a familiaridade com alguma linguagem de programação (sugere-se o software *Mathematica*).

## Conteúdo



[1] http://en.wikipedia.org/wiki/Numerical_analysis

# 1 Erros inerentes à representação numérica

Com o objectivo de instruirmos uma máquina computacional a realizar rapidamente muitos cálculos básicos, temos em primeiro lugar de definir uma representação dos elementos de $\mathbb{R}$. A primeira restrição óbvia tem a ver com os números arbitrariamente grandes. Outra restrição deriva da natureza aritmética dos números. Por exemplo, os números irracionais em $\mathbb{R} \setminus \mathbb{Q}$ são os que se escrevem em dízimas infinitas. Qualquer truncação desta dízima corresponde a um número racional. Ora, a capacidade de memória de um qualquer computador, quer seja agora no século XXI ou no XXX, será sempre finita. Logo, não nos é possível manipular um número irracional tendo em conta a sua expressão numérica.

Assim, o conjunto dos números representáveis numericamente constituem um subconjunto finito de $\mathbb{Q}$. Este subconjunto será o objecto do nosso estudo, tendo em conta a escolha inicial de uma forma consistente de representação de números.

## 1.1 Representação de $\mathbb{Z}$

Apesar de podermos sempre considerar representações de números reais em qualquer base, usamos frequentemente a base decimal[2]. Note que todas as representações em bases diferentes são equivalentes entre si, como iremos ver em seguida. Começamos pelo caso de números inteiros, i.e. o conjunto $\mathbb{Z}$. De facto, basta concentrarmo-nos em $\mathbb{N}$ pois zero é igual em qualquer base e para obtermos números negativos basta acrescentar em frente ao número o sinal $-$.

Considere uma base arbitrária $b \in \{2, 3, 4, \dots\}$ e o conjunto $\mathcal{A}_b$ de dígitos permitidos nessa base, i.e.

$$\#\mathcal{A}_b = b.$$

Podemos escolher os símbolos que quiseremos como elementos de $\mathcal{A}_b$ desde que sejam distintos entre si e totalizem $b$. Obviamente que os mais utilizados são os algarismos árabes. Por exemplo, podemos fazer corresponder à base binária ($b = 2$) os conjuntos $\mathcal{A}_2 = \{0, 1\}$, $\mathcal{A}_2 = \{V, F\}$, $\mathcal{A}_2 = \{H, M\}$, etc.

Sejam $a_i$ os elementos de $\mathcal{A}_b$ onde $i$ pertence a um subconjunto de $\mathbb{N}$. Escrevemos um número inteiro na base $b$ na forma $(a_n \dots a_0)_b$ para algum $n \in \mathbb{N}$. No caso da base decimal, i.e. $b = 10$, escrevemos simplesmente $a_n \dots a_0$, por ser esta a base usual.

[2]Provavelmente relacionado com o facto dos humanos terem 10 dedos nas mãos.

A transformação para a base decimal é dada por

$$(a_n \dots a_0)_b = \sum_{i=0}^{n} a_i b^i. \tag{1.1}$$

Quando não se especifica o conjuntos dos dígitos para uma base $b$, consideramos $\mathcal{A}_b = \{0, \dots, b-1\}$.

**Exemplo 1.1.** Confira:

1. $(1010)_2 = 10$
2. $(813)_9 = (220110)_3 = 660$
3. $(864D5F9A)_{16} = (10000110010011010101111110011010)_2 = 2253217690$ onde usámos

$$\mathcal{A}_{16} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}.$$

**Exercício 1.2.** *Aritmética em base b:*

1. *Descreva um algoritmo para a soma de dois números inteiros em base binária.*
2. *Desenvolva a forma de fazer o mesmo para uma base arbitrária b.*
3. *Repita as alíneas anteriores para a multiplicação.*

A relação entre as representações em bases genéricas $b$ e $b'$ é dada por

$$(a_n \dots a_0)_b = \left( \sum_{i=0}^{n} \alpha_i \beta^i \right)_{b'},$$

onde $\alpha_i$ e $\beta$ são as representações de $a_i$ e $b$ na base $b'$, respectivamente.

**Exemplo 1.3.** Queremos determinar $(a_2 a_1 a_0)_b = (813)_9$ na base $b' = 3$. Ora, $a_2 = (22)_3$, $a_1 = (1)_3$, $a_0 = (10)_3$ e $b = (100)_3$. Assim,

$$(813)_9 = (22)_3 (100)_3^2 + (1)_3 (100)_3 + (10)_3 = (220000)_3 + (100)_3 + (10)_3 = (220110)_3.$$

Um algoritmo simples para a representação de $x \in \mathbb{N}$ (base decimal) em base $b$ é o seguinte. Queremos determinar os $a_i$ tais que $x = (a_n \dots a_0)_b$. Assim, o resto da divisão inteira de

$$x = \sum_{i=0}^{n} a_i b^i = a_n b^n + \dots + a_1 b + a_0,$$

por $b$ é igual a $a_0$ pois

$$\frac{x}{b} = a_n b^{n-1} + \dots + a_1 + \frac{a_0}{b}.$$

Em particular,

$$a_0 = b\left(\frac{x}{b} - \left[\frac{x}{b}\right]\right)$$

onde $[y]$ é a parte inteira do número $y$. Finalmente, obtemos todos os dígitos $a_i$ usando a seguinte fórmula recursiva:

$$x_0 = x, \quad x_{i+1} = \left[\frac{x_i}{b}\right], \quad a_i = b\left(\frac{x_i}{b} - \left[\frac{x_i}{b}\right]\right).$$

## 1.2 Representação de $\mathbb{R}$

Como já sabemos mudar bases de representação de $\mathbb{Z}$, para completar $\mathbb{R}$ falta-nos considerar números em $]0,1[$. De facto, podemos sempre escrever

$$x = (a_n \dots a_0.a_{-1} \dots a_{-m})_b \in \mathbb{R}$$

como a soma da sua parte inteira $[x] = (a_n \dots a_1 a_0)_b$ com a sua parte fraccionária $x - [x] = (0.a_{-1}a_{-2} \dots a_{-m})_b$. Note que aqui $m \in \mathbb{N} \cup \{\infty\}$. A parte fraccionária na representação decimal é então dada por $\sum_{i=-m}^{-1} a_i b_i$. Finalmente,

$$x = \sum_{i=-m}^{n} a_i b^i.$$

**Exemplo 1.4.**

1. $(101.1001)_2 = 5.5625$
2. $(0.22222\dots)_3 = 1$

**Exercício 1.5.**

1. *Use as ideias da secção anterior para obter uma mudança de base da parte fraccionária em representação decimal para uma qualquer base $b$.*
2. *Encontre um algoritmo de mudança de base entre quaisquer duas bases.*
3. *Encontre algoritmos para a subtracção e divisão de números numa base arbitrária $b$.*
4. *Será que os irracionais em base $b$ também correspondem a partes fraccionárias infinitas não periódicas?*

**Exercício 1.6.**

1. *Indique a representação na base 2 dos números $(C1DADE.DE)_{16}$ e $(715B0A)_{16}$.*
2. *Mostre que para qualquer base $b \geq 2$ e $-n \leq i \leq m$,*

$$(10)_b^i (a_n \dots a_0.a_{-1} \dots a_{-m})_b = (a_n \dots a_{-i}.a_{-i-1} \dots a_{-m})_b.$$

## 1.3 Representação em ponto flutuante

A representação em "ponto flutuante"[3] é a mais apropriada para uso computacional. Fixando uma base $b$, podemos escrever um número $x \in \mathbb{R} \setminus \{0\}$ na forma

$$x = \pm(0.a_1a_2\dots)_b \times (10)_b^t$$

onde $a_1 \neq 0$ e $t \in \mathbb{Z}$. A representação em ponto flutuante de $x$ é assim dada pelo sinal $\pm$, a mantissa $m = a_1a_2\dots$, a base $b$ e o expoente $t$:

$$(\pm, m, b, t)$$

O número zero é um caso especial, sendo representado simplesmente por 0.

**Exemplo 1.7.** Seja $x = -0.00012$. Temos assim que $x = -0.12 \times 10^{-3}$ na base decimal, e $x = -(0.111111)_2 \times (10)_2^{-13}$ na base binária.

Como já foi referido, um computador apenas trabalha com um número finito de dígitos. Assim, com essa restrição, temos que "arrendondar" o valor de $m$ e definir limites para o expoente $t$ na expressão acima para $\widetilde{m}$ e $\widetilde{t}$, respectivamente. Considere $p \in \mathbb{N}$ como o número máximo de dígitos da mantissa (precisão). Assim,

$$\widetilde{m} = 0.\widetilde{a}_1 \dots \widetilde{a}_p$$

onde $\widetilde{a}_1 \neq 0$. Adicionalmente, considere $q, q' \in \mathbb{N}$ como os valores limites de $-q' \leq \widetilde{t} \leq q$.

A regra de arredondamento que iremos respeitar é a da melhor aproximação. Em caso de ambiguidade arredondamos o último dígito para o caso par (outras opções são igualmente válidas).

**Exemplo 1.8.** Sejam $b = 10$, $p = 2$ e $q = q' = 4$. Note que qualquer número no intervalo $[0.00012, 0.000125]$ tem a mesma representação $0.12 \times 10^{-3}$.

Com as condições acima impostas, os valores de $\widetilde{m}$ e de $\widetilde{t}$ ficam restringidos a:

$$\frac{1}{b} \leq \widetilde{m} \leq 1 - \frac{1}{b^p}, \qquad \widetilde{t} \in \{-q', \dots, q\}.$$

Defina o conjunto $Q_{b,p,q,q'}$ dos números representáveis em ponto flutuante com as restrições acima impostas, incluindo o zero.

**Exercício 1.9.**

1. *Prove que:*

   (a) $\#Q_{b,p,q,q'} = 2(b-1)b^{p-1}(q+q'+1)+1$

[3] ou vírgula flutuante conforme a notação usada para a separação entre a parte inteira e a parte fraccionária.

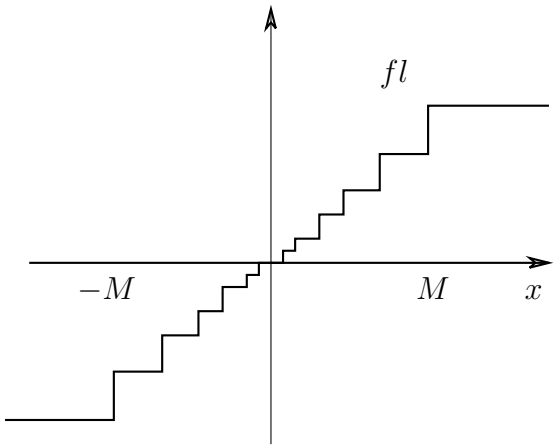


Figura 1: Exemplo de função $fl$.

*(b)* $M := \max Q_{b,p,q,q'} = (1 - b^{-p})b^q = -\min Q_{b,p,q,q'}$

*(c)* $\min Q_{b,p,q,q'} \cap \mathbb{R}^+ = b^{-(q'+1)}$.

2. *Conclua que $Q_{b,p,q,q'}$ é um conjunto finito e que*

$$\mathbb{Z} \cap [-M, M] \subset Q_{b,p,q,q'} \subset \mathbb{Q} \subset \mathbb{R}.$$

**Exemplo 1.10.** A norma IEEE754, a mais comum nos computadores actuais a 64 bits) estabelece que um número com precisão dupla corresponde a $b = 2$, $p = 52$, $q = 1024$ e $q' = q - 1$. O vector $(\pm, m, t)$ é assim guardado em 64 dígitos binários (bits): um para o sinal, 52 para a mantissa, 10 para o valor máximo do expoente e um para o sinal do expoente. Temos assim que o número total de números representados nesta norma é de $0.922337 \times 10^{19}$, sendo o valor máximo $0.179769 \times 10^{309}$ e o mínimo positivo $0.5562684 \times 10^{-308}$ (valores aproximados).

Podemos agora definir uma função que a dado número real obtemos a sua representação numérica, i.e. a melhor aproximação em representação de ponto flutuante. Seja a função

$$fl \colon \mathbb{R} \to Q_{b,p,q,q'}$$

tal que

$$fl(x) - x = \min_{z \in Q_{b,p,q,q'}} |z - x|$$

e em caso de ambiguidade escolhemos o valor de $fl(x)$ onde o último dígito é par (ver Figura 1).

Considere $x = \pm m b^t$ dentro dos limites de representação dados por $[-M, M]$. Assim,

$$fl(x) = \widetilde{m} b^t.$$

**Exemplo 1.11.** Sejam $b = 10$, $p = 5$, $q = 10$ e $q' = 9$.

1. $fl(327) = 0.327 \times 10^3$
2. $fl(\pi) = 0.31416 \times 10^1$

## 1.4 Erros de representação

O erro absoluto de aproximação é dado por

$$\begin{aligned} E(x) &= fl(x) - x \\ &= (\widetilde{m} - m)b^t. \end{aligned}$$

Esta grandeza não é contudo de grande utilidade. Se estivermos a trabalhar com valores muito pequenos, então $E(x)$ será sempre muito baixo. O que interessa é a relação entre o erro absoluto e os valores que estamos a considerar. Assim, definimos o **erro relativo** como

$$e(x) = \frac{E(x)}{x}$$

Note que podemos agora escrever

$$fl(x) = x[1 + e(x)].$$

**Exercício 1.12.** *Para $x \in [-M, M]$, prove as seguintes estimativas por cima dos erros:*

1. $|E(x)| \leq \frac{1}{2}b^{t-p}$
2. $|e(x)| \leq \frac{1}{2}b^{1-p}$

## 1.5 Erros nas operações aritméticas

### 1.5.1 Soma

Ao somarmos computacionalmente dois números reais através das suas aproximações, a soma também será uma aproximação do valor exacto. Queremos estimar o erro de representação associado a esta operação.

Tendo $x_i = m_i b^{t_i}$, $i = 1, 2$, e incluindo o sinal de $x_i$ em $m_i$, obtemos

$$x_1 + x_2 = \begin{cases} [m_1 + m_2 b^{-(t_1 - t_2)}]b^{t_1}, & t_1 > t_2 \\ [m_1 b^{-(t_2 - t_1)} + m_2]b^{t_2}, & t_2 \geq t_1. \end{cases}$$

Assim, relembrando que $fl(x_i) = x_i(1 + \varepsilon_i)$ com $\varepsilon_i = e(x_i)$,

$$fl(x_1) + fl(x_2) = x_1(1 + \varepsilon_1) + x_2(1 + \varepsilon_2).$$

Para este valor ser representado numericamente é necessário aplicar novamente a função $fl$. Então,

$$\begin{aligned} fl(fl(x_1)+fl(x_2)) &= [x_1(1+\varepsilon_1)+x_2(1+\varepsilon_2)]\,(1+\sigma) \\ &= x_1(1+\varepsilon_1)(1+\sigma)+x_2(1+\varepsilon_2)(1+\sigma), \end{aligned}$$

onde $\sigma = e(fl(x_1)+fl(x_2))$. Logo, o erro absoluto é dado por

$$fl(fl(x_1)+fl(x_2))-(x_1+x_2) = x_1\varepsilon_1(1+\sigma)+x_2\varepsilon_2(1+\sigma)+(x_1+x_2)\sigma.$$

**Exemplo 1.13.** Sejam $b=10$, $p=4$, $q=10$, $x_1 = 0.43787\times 10^{-2}$ e $x_2 = 0.43783\times 10^{-2}$. Queremos calcular a representação aproximada de $x_1 - x_2$:

$$fl(fl(x_1)-fl(x_2)) = fl(0.0001\times 10^{-2}) = 0.1\times 10^{-5}.$$

No entanto $x_1 - x_2 = 0.4\times 10^{-6}$. Podemos pensar que o módulo do erro absoluto $|E| = 0.6\times 10^{-6}$ é um valor pequeno, mas é superior ao valor exacto da subtracção! O módulo do erro relativo $|e| = 1.5$ já nos mostra que o erro é relevante.

**Observação 1.14.** Como vimos no exemplo anterior, quando se subtraem números muito próximos pode-se obter um erro relativo demasiado pessimista. Em alternativa para esta situação, para verificarmos o grau de viabilidade do resultado, podemos calcular $E/x_1$.

### 1.5.2 Somas finitas

Podemos agora representar numericamente uma soma finita $\sum_{i=1}^{n} x_i$. Para isso utilizamos uma algoritmo simples para o seu cálculo[4].

Considere $S_0 = 0$ e $S_i = S_{i-1}+x_i$ com $i = 1,\ldots,n$. Logo, indutivamente obtemos $S_n = \sum_{i=1}^{n} x_i$. A representação numérica é assim calculada por $\widetilde{S}_0 = 0$ e

$$\begin{aligned} \widetilde{S}_i &= fl\left(\widetilde{S}_{i-1}+fl(x_i)\right) \\ &= \left(\widetilde{S}_{i-1}+fl(x_i)\right)(1+\sigma_i), \end{aligned} \tag{1.2}$$

onde $\sigma_i$ é o erro relativo da soma $i$. O erro absoluto final é $E = \widetilde{S}_n - S_n$ e o relativo $e = E/S_n$.

**Exercício 1.15.** *Deduza uma fórmula explícita para a propagação de erros no cálculo de $\sum_i^n x_i$ usando o algoritmo $S_0 = 0$, $S_i = S_{i-1}+x_i$, $i = 1,\ldots,n$. Isto é, calcule $\widetilde{S}_n - S_n$ dependendo de $n$.*

[4]Podem existir outros eventualmente mais eficientes. Por exemplo, reordenando a soma de acordo com a ordem de grandeza dos termos.

### 1.5.3 Multiplicação

Sejam $x_1$ e $x_2$ para os quais temos $fl(x_i) = x_i(1+\varepsilon_i)$ com $\varepsilon_i = e(x_i)$. Queremos determinar a representação numérica do produto $x_1x_2$. Assim,

$$\begin{aligned} fl(fl(x_1)\, fl(x_2)) &= fl\,(x_1x_2(1+\varepsilon_1)(1+\varepsilon_2)) \\ &= x_1x_2(1+\varepsilon_1)(1+\varepsilon_2)(1+\mu), \end{aligned} \tag{1.3}$$

onde $\mu = e(fl(x_1)\, fl(x_2))$. Finalmente, o erro absoluto é dado por $E = x_1x_2[(1+\varepsilon_1)(1+\varepsilon_2)(1+\mu)-1]$ e o relativo $e = (1+\varepsilon_1)(1+\varepsilon_2)(1+\mu)-1$.

**Exercício 1.16.** *Deduza uma fórmula explícita para a propagação de erros no cálculo do produto interno usual* $\langle x, y\rangle = \sum_{i=1}^{n} x_iy_i$ *usando o algoritmo* $S_0 = 0$, $S_i = S_{i-1} + x_iy_i$, $i = 1, \ldots, n$.

**Exercício 1.17.** *Escreva as fórmulas dos erros numéricos absoluto e relativo para a divisão* $x_1/x_2$ *com* $x_2 \neq 0$.

### 1.5.4 Funções $C^1$

O caso geral de uma função continuamente diferenciável é também simples de tratar. Seja $f \in C^1(D, \mathbb{R})$ com $D \subset \mathbb{R}$ e $fl(D) \subset D$. Para representarmos esta função numericamente podemos ter que definir uma aproximação numérica da função. Dado $\epsilon > 0$, esta será uma nova função dada por

$$\widetilde{f}\colon fl(D) \to \mathbb{R}$$

tal que

$$\sup_{z \in fl(D)} |\widetilde{f}(z) - f(z)| < \epsilon.$$

Queremos obviamente considerar $\epsilon$ tão pequeno quanto possível, podendo mesmo ser que $\widetilde{f} = f$ (se $f$ for uma função mal conhecida, podemos aproximá-la por exemplo por um polinómio).

A representação numérica da função $f$ será então dada pela transformação (cf. Figura 2)

$$x \mapsto fl \circ \widetilde{f} \circ fl(x).$$

O erro cometido pode ser estimado por

$$\begin{aligned} E_f(x) &:= fl \circ \widetilde{f} \circ fl(x) - f(x) \\ &= fl \circ \widetilde{f} \circ fl(x) - \widetilde{f} \circ fl(x) + (\widetilde{f} - f) \circ fl(x) + f \circ fl(x) - f(x) \\ &= E(\widetilde{f} \circ fl(x)) + (\widetilde{f} - f) \circ fl(x) + f'(\xi)[fl(x) - x] \end{aligned}$$

com $\xi$ entre $x$ e $fl(x)$.

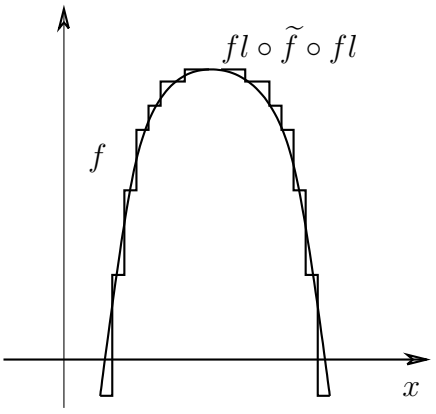


Figura 2: Exemplo de representação numérica da função $f$.

**Exercício 1.18.** *Mostre que o erro relativo dado por $e_f(x) = E_f(x)/f(x)$ é aproximadamente*

$$e_f(x) \simeq \frac{xf'(x)}{f(x)} e(x).$$

*Ou seja, o termo $xf'(x)/f(x)$ estima a propagação dos erros.*

**Exemplo 1.19.** Assuma os parâmetros $b = 10$, $p = 6$ e $q = 99$ para a representação numérica de ponto flutuante. Para as funções seguintes consideramos aproximações $\widetilde{f}$ tais que $|(\widetilde{f} - f) \circ fl(x)| < 0.5 \times 10^{-5}$.

1. $f(x) = \sqrt{x}$, $x > 0$. Fazendo $\xi \geq x - |fl(x) - x| = x - |xe(x)| \geq x(1 - 0.5 \times 10^{-5})$ obtemos

$$\begin{aligned} |e_f(x)| &\leq 0.1 \times 10^{-4} + \frac{x}{2\sqrt{\xi x}} 0.5 \times 10^{-5} \\ &\leq 0.1 \times 10^{-4} + 0.6703163715 \times 10^{-8}, \end{aligned}$$

   que não depende de $x$ e é "pequeno". Note que $|xf'(x)/f(x)| = 1/2$.

2. $f(x) = x^2$. Então, usando $\xi \leq x + |fl(x) - x| \leq x(1 + 0.5 \times 10^{-5})$,

$$\begin{aligned} |e_f(x)| &\leq 0.1 \times 10^{-4} + \frac{2\xi x}{x^2} 0.5 \times 10^{-5} \\ &\leq 0.1 \times 10^{-4} + 0.1 \times 10^{-4}. \end{aligned}$$

   Note que $|xf'(x)/f(x)| = 2$.

3. $f(x) = e^x$. Então,

$$\begin{aligned} |e_f(x)| &\leq 0.1 \times 10^{-4} + \frac{\mathrm{e}^{\xi} x}{e^x} 0.5 \times 10^{-5} \\ &\leq 0.1 \times 10^{-4} + 0.5 \times 10^{-5} e^{0.5 \times 10^{-5} x} x, \end{aligned}$$

dependendo de $x$, logo podendo ser muito "grande". Note que $|xf'(x)/f(x)| = |x|$.

**Exercício 1.20.** *Encontre fórmulas dos erros para uma função de duas variáveis $\varphi \in C^1(D, \mathbb{R})$ onde $D \subset \mathbb{R}^2$ e $Fl(D) \subset D$ com $Fl(x, y) = (fl(x), fl(y))$. Obtenha o resultado do exercício 1.17 como caso particular.*

# 2 Métodos numéricos para equações lineares

Nesta secção iremos estudar métodos para determinar a solução $x \in \mathbb{R}^d$ de um sistema de equações lineares na forma

$$Ax = b,$$

onde $A$ é uma matriz[5] real $d \times d$ invertível ($\det A \neq 0$) e $b \in \mathbb{R}^d$. Isto equivale a calcular a inversa de $A$ e a multiplicá-la por $b$. O método de inversão de matrizes mais aconselhável na perspectiva de um analista numérico é aquele que optimiza o tempo de cálculo e/ou minimiza o erro.

## 2.1 Matrizes triangulares

Começamos com o caso particular das matrizes triangulares superiores. Ou seja, matrizes $A = [a_{ij}]$ tais que $a_{ij} = 0$ se $i > j$. Recorde que $\det A \neq 0$, logo $a_{ii} \neq 0$ para qualquer $0 \leq i \leq d$.

**Exercício 2.1.**

1. *Mostre que a solução de $Ax = b$ com $A$ triangular superior invertível é dada por*

$$x_d = \frac{b_d}{a_{dd}}, \qquad x_i = \frac{b_i - \sum_{j=i+1}^{d} a_{ij} x_j}{a_{ii}}, \quad i = d-1, \ldots, 1.$$

2. *Justifique que o número de operações aritméticas efectuadas para calcular a solução acima é dado por:*
   - *Número de divisões: $d$*
   - *Número de multiplicações: $d(d-1)/2$*
   - *Número de somas: $d(d-1)/2$*
   - *Número total de operações: $d^2$*
3. *Repita o procedimento para matrizes triangulares inferiores (matrizes transpostas de triangulares superiores).*

[5] O espaço das matrizes $d \times d$ com entradas reais e invertíveis é denotado por $\mathcal{M}_{d\times d}(\mathbb{R})$.

## 2.2 Matrizes não triangulares

### 2.2.1 Método de Gauss

O caso geral de matrizes não triangulares é mais complicado. Envolve recorrer ao método de Gauss para obtermos um sistema "triangular", como anteriormente.

**Exemplo 2.2.** Considere a equação linear $Ax = b$ onde

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

O 1º passo do método de Gauss corresponde à eliminação das entradas na coluna por debaixo do pivot $a_{11} = 1$:

$$\left[\begin{array}{ccc|c} 1 & 0 & 2 & 1 \\ 2 & 2 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 2 & 1 \\ 0 & 2 & -3 & -2 \\ 0 & 1 & -1 & -1 \end{array}\right].$$

Ou seja, a nova 2ª linha obteve-se pela multiplicação da 1ª linha por $-a_{21}/a_{11} = -2$ somando à 2ª, e a nova 3ª linha pela multiplicação da 1ª por $-a_{31}/a_{11} = -1$ somando à 3ª. Este passo pode ser resumido como a multiplicação pela matriz

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

dando origem a uma equação equivalente

$$M_1 A x = M_1 b.$$

O 2º passo envolve a escolha de pivot da entrada em $(2, 2)$, que denotamos por $b_{22}$, e a eliminação da entrada por debaixo.

$$\left[\begin{array}{ccc|c} 1 & 0 & 2 & 1 \\ 0 & 2 & -3 & -2 \\ 0 & 1 & -1 & -1 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 2 & 1 \\ 0 & 2 & -3 & -2 \\ 0 & 0 & 1/2 & 0 \end{array}\right]$$

Para a obtenção da nova 3ª linha multiplicámos por $-1/2$ a 2ª linha que somámos à 3ª. Como anteriormente, este passo resume-se à multiplicação por

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/2 & 1 \end{bmatrix}$$

dando origem a uma equação equivalente

$$M_2 M_1 A x = M_2 M_1 b.$$

Ora, $U = M_2 M_1 A$ é uma matriz triangular superior. Temos assim uma equação linear triangular, de fácil resolução. Isto é,

$$x = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

**Exercício 2.3.** *Baseando-se no exemplo anterior (onde não foi necessário haver trocas de linhas de forma a obtermos pivots diferentes de zero), confira o número de operações aritméticas efectuadas para uma matriz $d \times d$ são dados por:*

- *$1^o$ passo do método de Gauss:*
  - *Número de divisões:* $d - 1$
  - *Número de multiplicações:* $d(d-1)$
  - *Número de somas:* $d(d-1)$
  - *Número total:* $(d-1)(2d+1)$
- *Totais do método de Gauss:*
  - *Número de divisões:* $d(d-1)/2$
  - *Número de multiplicações:* $d(d^2-1)/3$
  - *Número de somas:* $d(d^2-1)/3$
  - *Número total:* $d(d-1)\left(\frac{2}{3}d + \frac{7}{6}\right)$

**Observação 2.4.** Note que o cálculo de inversas de matrizes pela Regra de Cramer envolve $n!$ parcelas e $n!(n-1)$ multiplicações decorrentes da necessidade de calcular determinantes de matrizes. O número de operações aritméticas necessárias torna este método inviável do ponto de vista numérico para valores de $n$ acima de 10.

### 2.2.2 Pesquisas de pivot

Por simplicidade começámos por estudar um caso onde pudemos escolher pivots diferentes de zero ao longo de todos os passos do método de Gauss, sem necessidade de troca de linhas. O exemplo seguinte ilustra o caso onde esse facto pode levar a erros numéricos consideráveis.

**Exemplo 2.5.** Considere a equação linear $Ax = b$ com

$$A = \begin{bmatrix} 10^{-6} & 1 \\ 1 & 1 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}.$$

Queremos resolver este sistema por diferentes variações do método de Gauss.

1. Usando o pivot $a_{11} = 10^{-6}$:

$$\left[\begin{array}{cc|c} 10^{-6} & 1 & 0.5 \\ 1 & 1 & 1 \end{array}\right] \to \left[\begin{array}{cc|c} 10^{-6} & 1 & 0.5 \\ 0 & 1-10^{6} & 1-0.5\times 10^{6} \end{array}\right].$$

   A solução exacta é então

$$x = \begin{bmatrix} 0.5000005 \\ 0.4999995 \end{bmatrix}.$$

2. Usando o pivot $a_{11} = 10^{-6}$ e tendo em conta as representações numéricas em ponto flutuante com máximo de 6 dígitos na parte decimal:

$$\left[\begin{array}{cc|c} 10^{-6} & 1 & 0.5 \\ 1 & 1 & 1 \end{array}\right] \to \left[\begin{array}{cc|c} 0.1\times 10^{-5} & 0.1\times 10^{1} & 0.5 \\ 0 & -0.999999\times 10^{6} & -0.499999\times 10^{6} \end{array}\right].$$

   Logo, a solução numérica aproximada é

$$\widetilde{x} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}.$$

   A primeira componente de $\widetilde{x}$ difere completamente do valor exacto $x$ obtido anteriormente. O problema reside no uso de um pivot com um valor absoluto muito inferior às restantes entradas da matriz.

3. Usando **pesquisa parcial de pivot**. De forma a evitarmos erros como o anterior, podemos escolher para pivot de uma determinada coluna, a entrada com maior valor absoluto. Para isso basta-nos trocar as linhas:

$$\left[\begin{array}{cc|c} 10^{-6} & 1 & 0.5 \\ 1 & 1 & 1 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 1 & 1 \\ 10^{-6} & 1 & 0.5 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 1 & 1 \\ 0 & 1-10^{-6} & 0.5-10^{-6} \end{array}\right].$$

   Logo,

$$\widetilde{x} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

4. Uma forma de optimizar ainda mais o método de Gauss, usamos uma **pesquisa total de pivot**. Escolhemos para pivot a entrada de toda a matriz (não somente da coluna respectiva como na pesquisa parcial) com maior valor absoluto. Agora temos que trocar linhas e colunas. A troca de colunas implica uma troca da ordem das componentes da solução, que se terá que ter em conta no final:

**Exercício 2.6.** *Implemente computacionalmente o método de Gauss na sua linguagem preferida. Use o código para determinar a solução de $Ax = b$ com*

$A = [a_{ij}]_{i,j=1,\ldots,10}$ *e* $b \in \mathbb{R}^{10}$ *dados por*

$$a_{ij} = \begin{cases} 2, & i = j \\ -1, & |i-j| = 1 \\ 0, & o.c. \end{cases} \quad e \quad b = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

### 2.2.3 Factorização triangular

Como vimos, podemos reduzir a solução de uma equação linear ao caso "triangular" usando uma factorização triangular obtida pelo método de Gauss. Em muitas aplicações é necessário calcular soluções da equação linear $Ax = b$ para vários valores de $b$, mantendo $A$. De forma a evitar ter que resolver numericamente o método de Gauss para cada um separadamente (usando cerca de $d^3$ operações), podemos memorizar a factorização e aplicá-la de forma simples (em $d^2$ operações) para cada $b$ dado. Isto é, no caso mais geral queremos escrever

$$PAQ = LU$$

com $L = [\ell_{ij}]$ triangular inferior e diagonal unitária ($\ell_{ii} = 1$) e $U = [u_{ij}]$ triangular superior. A matriz $P$ é uma permutação de linhas e $Q$ permuta colunas. O sistema $Ax = b$ pode assim ser escrito como $PAQz = L(Uz) = Pb$ onde $x = Qz$, dando origem a dois problemas "triangulares" de resolução simples:

$$Ly = Pb \quad \text{e} \quad Uz = y.$$

A solução é então obtida por $x = Qz$.

Note que a matriz $P$ corresponde ao método de pesquisa parcial de pivot referido anteriormente. A pesquisa total de pivot corresponde à utilização simultaneamente de $P$ e de $Q$, i.e. permutações de linhas e de colunas.

No Exemplo 2.2 já vimos como o método de Gauss nos dá uma factorização triangular. Nesse caso obtivemos $U = M_2 M_1 A$ onde as matrizes $M_i$ correspondem aos passos do método, e são matrizes triangulares inferiores com diagonal unitária. As suas inversas são ainda do mesmo tipo, assim como o produto. Logo, neste caso $L = M_1^{-1} M_2^{-1}$.

O caso de uma matriz geral invertível é similar. No $k^{\text{o}}$ passo (de entre $d-1$ passos totais), o método de Gauss corresponde a multiplicar pela matriz

$$M_k = [m_{ij}^{(k)}] \quad \text{onde} \quad m_{ij}^{(k)} = \begin{cases} 0, & i < j \\ 1, & i = j \\ \gamma_i^{(k)}, & j = k,\ i = k+1, \ldots, d \\ 0, & \text{o.c.} \end{cases}$$

Os números $\gamma_i^{(k)}$ são os multiplicadores (relacionados com o pivot e a entrada a eliminar) pela linha $k$ (a do pivot) que será somada à linha $i$. Se for necessário trocar primeiro de linhas pela pesquisa parcial de pivot, construímos a matriz de permutação entre as linhas $k$ e $m > k$:

$$P_k = [p_{ij}^{(k)}] \quad \text{onde} \quad p_{ij}^{(k)} = \begin{cases} 1, & i = j \quad \text{e} \quad i \neq m \quad \text{e} \quad i \neq k \\ 1, & i = m,\, j = k \\ 1, & i = k,\, j = m \\ 0, & \text{o.c.} \end{cases}$$

Note ainda que $P_k^2 = I$, $P_k^{-1} = P_k$ e $P_k^T = P_k$.

**Exercício 2.7.**

1. *Mostre que $M_k$ é invertível e calcule a sua inversa.*

2. *Calcule $P_{k+1}M_k^{-1}P_{k+1}$ e mostre que é uma matriz triangular inferior com diagonal unitária onde apenas a coluna k tem entradas abaixo da diagonal diferentes de zero. Repita para as matrizes $N_k = P_{d-1}\dots P_{k+1}M_k^{-1}P_{k+1}\dots P_{d-1}$ (considere $N_{d-1} = M_{d-1}^{-1}$).*

3. *Obtenha a matriz de permutações $P$ não trivial tal que*

$$PP_1M_1^{-1}\dots P_{d-1}M_{d-1}^{-1} = N_1\dots N_{d-1}.$$

Finalmente,

$$U = M_{d-1}P_{d-1}\dots M_1P_1A$$

é triangular superior e $L = PP_1M_1^{-1}\dots P_{d-1}M_{d-1}^{-1}$ é triangular inferior com diagonal unitária e $P = P_{d-1}\dots P_1$ pelo exercício acima. Logo, $PA = LU$.

**Exercício 2.8.** *Determine a decomposição $PA = LU$ da matriz $A$ usando o método de Gauss com pesquisa parcial de pivot, onde*

$$A = \begin{bmatrix} 1.4 & 1.42 & 6.5 \\ 2 & 1 & 1 \\ 0.4 & 1.4 & 3.2 \end{bmatrix}.$$

**Exercício 2.9.** *Repita o procedimento anterior considerando a cada passo do método de Gauss pesquisa total de pivots. Obtenha assim simultaneamente uma matriz de permutação de colunas $Q$ (não trivial) e prove a decomposição $PAQ = LU$.*

**Exercício 2.10.**

1. *A partir da factorização triangular de uma matriz $A$ invertível, determine um algoritmo para o cálculo de* $\det A$.

2. *Considere a representação numérica com parâmetros $p = 6$ e $q = 99$. Dê exemplo de uma matriz para a qual testar a sua invertibilidade pelo cálculo numérico do seu determinante daria um resultado errado.*

**Exercício 2.11** (Método de Doolitle)**.** *Considere as matrizes $d \times d$ dadas por $A = [a_{ij}]$, $U = [u_{ij}]$ e $L = [\ell_{ij}]$ tais que $A = LU$ é uma factorização triangular, $i = 1, \ldots, d$. Mostre que:*

1. *A primeira linha de $U$ é dada por $u_{1j} = a_{1j}$ e a primeira coluna de $L$ é $\ell_{i,1} = a_{i,1}/a_{1,1}$.*

2. *As entradas das matrizes $U$ e $L$ são dadas por*

$$u_{ij} = a_{ij} - \sum_{m=1}^{i-1} \ell_{im} u_{mj}, \quad j \geq i$$
$$\ell_{ij} = \frac{a_{ij} - \sum_{m=1}^{j-1} \ell_{im} u_{mj}}{u_{jj}}, \quad j < i.$$

## 2.3 Análise de erros

De forma a estimar erros associados aos vectores solução de uma equação linear, vamos primeiro estudar normas de vectores e de matrizes.

### 2.3.1 Normas de vectores e matrizes

Seja $E$ um espaço vectorial real. Uma norma em $E$ é uma aplicação $\|\cdot\| \colon E \to \mathbb{R}_0^+$ tal que

- $\|x\| = 0$ sse $x = 0$ (não degeneração),
- $\|\alpha x\| = |\alpha| \, \|x\|$, $x \in E$ e $\alpha \in \mathbb{R}$ (linearidade),
- $\|x + y\| \leq \|x\| + \|y\|$, $x, y \in E$ (desiguldade triangular).

**Exemplo 2.12.** Considere $E = \mathbb{R}^d$ e $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

1. As normas $\ell_p$, $p \in \mathbb{N}$, são dadas por

$$\|x\|_p = \left[ \sum_{i=1}^{d} |x_i|^p \right]^{1/p}.$$

   As mais usadas são $\ell_1$ e $\ell_2$ (conhecida por euclideana).

2. A norma $\ell_\infty$ é

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

Note que a implementação numérica das diferentes normas anteriores varia conforme o número de operações aritméticas necessárias para o seu cálculo, e as estimativas dos erros associados.

**Exercício 2.13.** *Esboce as circunferências centradas na origem e raio $r > 0$ em $\mathbb{R}^2$, definidas por $\{x \in \mathbb{R}^2 \colon \|x\|_p = r\}$, para $p = 1, 2, \infty$.*

**Exemplo 2.14.** Seja $E = \mathcal{M}_{d\times d}(\mathbb{R})$ e $A = [a_{ij}]_{i,j=1,\dots,d} \in \mathcal{M}_{d\times d}(\mathbb{R})$.

1. Normas induzidas pelo isomorfismo entre espaços lineares com a mesma dimensão finita $\mathcal{M}_{d\times d}(\mathbb{R}) \to \mathbb{R}^d$, e.g. $\|A\| = \sum_{i,j} |a_{ij}|$.

2. $\|A\|_1 = \max_j \|Ae_j\|_1 = \max_{j=1,\dots,d} \sum_{i=1}^d |a_{ij}|$ corresponde ao máximo das normas $\ell_1$ dos vectores coluna.

3. $\|A\|_\infty = \max_i \|e_i^T A\|_1 = \max_{i=1,\dots,d} \sum_{j=1}^d |a_{ij}|$ corresponde ao máximo das normas $\ell_1$ dos vectores linha.

4. $\|A\|_2 = \sqrt{\rho(A^T A)}$ onde $\rho(B)$ é o raio espectral, i.e. o máximo valor próprio de $B$ em valor absoluto. Note que $A^T A$ é uma matriz simétrica, logo pode ser diagonalizável por matrizes ortogonais[6] e os valores próprios são reais e não negativos.

5. Sendo
$$A = \begin{bmatrix} 1 & 2 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$
obtemos $\|A\|_1 = 4$, $\|A\|_\infty = 6$ e $\|A\|_2 = 3.86432...$

Duas normas $\|\cdot\|$ e $\|\cdot\|'$ em $E$ são equivalentes sse existe $C > 1$ tal que para qualquer $x \in E$ temos

$$C^{-1}\|x\| \leq \|x\|' \leq C\|x\|.$$

Ou seja, podemos sempre comparar as normas de um vector $x$ uniformemente (usando a mesma constante.

**Teorema 2.15.** *Todas as normas para um espaço de dimensão finita são equivalentes.*

*Demonstração.* Considere um espaço vectorial real $E$ com dimensão $d$. Seja $\{e_i\}_{i=1,\dots,d}$ uma base de $E$, tal que qualquer $x \in E$ pode ser escrito como $x = \sum_{i=1}^d x_i e_i$, onde $x_i \in \mathbb{R}$. É suficiente mostrar que qualquer norma $\|\cdot\|$ em $E$ é equivalente à norma $\|x\|_1 = \sum_{i=1}^d |x_i|$ uma vez que a equivalência de normas é transitiva. Para isso basta provar que para $\|x\|_1 = 1$ obtemos $C_1 \leq \|x\| \leq C_2$.

[6] $M$ é uma matriz ortogonal se $M^T M = MM^T = I$.

Note que $f\colon x \to \|x\|$ é uma função contínua em $E$ munido com a norma $\|\cdot\|_1$. Ou seja, dado $x_0 \in E$,

$$\lim_{\|x-x_0\|_1 \to 0} \|x\| = \|x_0\|.$$

De facto, dados $x, y \in E$ temos

$$\|x\|-\|y\| = \|y+(x-y)\|-\|y\| \leq \|x-y\| \leq \sum_i |x_i-y_i|\|e_i\| \leq \max_i \|e_i\| \, \|x-y\|_1.$$

Logo, $|\|x\| - \|x_0\|| \leq \max_i \|e_i\| \, \|x - x_0\|_1 \to 0$.

Deste modo, $f$ tem máximo e mínimo no compacto $\{x \in E\colon \|x\|_1 = 1\}$. Esses valores determinam assim a existência das constantes $C_1$ e $C_2$ acima. □

### 2.3.2 Normas de matrizes como operadores lineares

Sejam $(E, \|\cdot\|_E)$ e $(F, \|\cdot\|_F)$ espaços vectoriais normados. Definimos $\mathcal{L}(E, F)$ como o espaço das transformações (operadores) lineares $E \to F$.

A norma usual em $\mathcal{L}(E, F)$ induzida pelas normas em $E$ e $F$ é dada por:

$$\|A\|_{\mathcal{L}} = \sup_{x \in E \setminus \{0\}} \frac{\|Ax\|_F}{\|x\|_E} = \sup_{\|x\|_E = 1} \|Ax\|_F, \qquad A \in \mathcal{L}(E, F).$$

**Exercício 2.16.** *Demonstre que $\|\cdot\|_{\mathcal{L}}$ é uma norma.*

**Observação 2.17.** Observe que para qualquer $x \in E$ temos $\|Ax\|_F \leq \|A\|_{\mathcal{L}} \, \|x\|_E$.

O subconjunto de $\mathcal{L}(E, F)$ que corresponde aos operadores $A$ com norma limitada, $\|A\|_{\mathcal{L}} < \infty$, é denotado por $L(E, F)$. Para espaços de dimensão finita temos sempre $\mathcal{L}(E, F) = L(E, F)$ (basta pensar em termos de $\mathbb{R}^d$ pois todos os com a mesma dimensão finita são isomorfos).

**Observação 2.18.** Apesar de podermos definir normas para o espaço de matrizes como vimos na secção anterior, interessa-nos considerar normas matriciais induzidas por normas vectoriais interpretando as matrizes como operadores lineares. De facto, o nosso interesse centra-se na obtenção de soluções vectoriais de equações lineares, onde o papel da matriz é o do operador. Assim, nestas notas restringimo-nos a normas matriciais induzidas pelas normas vectoriais indicadas, e escrevemos $\|\cdot\| = \|\cdot\|_{\mathcal{L}}$.

**Exemplo 2.19.** Considere $E = F = \mathbb{R}^d$ munidos da norma $\ell_1$ e $L(\mathbb{R}^d, \mathbb{R}^d) = \mathcal{M}_{d\times d}(\mathbb{R})$. Seja $A = [a_{ij}]_{i,j} \in \mathcal{M}_{d\times d}(\mathbb{R})$ e $e_j$ os vectores da base canónica de $\mathbb{R}^d$. Em particular $\|e_j\|_1 = 1$. Assim,

$$\|A\| = \sup_{\|x\|_1 = 1} \|Ax\|_1 \geq \max_j \|Ae_j\|_1 = \|A\|_1.$$

Por outro lado, para $\|x\|_1 = 1$,

$$\|Ax\|_1 = \left\| \left[ \sum_j a_{ij} x_j \right]_i \right\|_1 = \sum_i \left| \sum_j a_{ij} x_j \right| \leq \sum_i \max_j |a_{ij}| \sum_j |x_j| = \|A\|_1$$

Isto implica que $\|A\| \leq \|A\|_1$. Finalmente,

$$\|A\|_1 \leq \|A\| \leq \|A\|_1,$$

logo $\|A\| = \|A\|_1$.

**Exercício 2.20.** *Repita o exemplo anterior para a norma $\ell_\infty$.*

**Exemplo 2.21.** Considere agora a norma $\ell_2$ em $\mathbb{R}^d$. Restriginmo-nos aqui ao caso de matrizes $A = [a_{ij}]_{i,j}$ pertencentes ao espaço $\mathcal{S}_{d\times d}(\mathbb{R})$ das matrizes simétricas $d \times d$ com coeficientes reais. A matriz simétrica $A^T A$ tem valores próprios reais $\lambda_i^2$ onde $\lambda_i$ são os valores próprios (reais) de $A$. Logo, $\|A\|_2 = \rho(A) = \max_i |\lambda_i|$. Os vectores próprios $v_i$ (normalizados $\|v_i\|_2 = 1$) da matriz simétrica formam uma base ortonormada de $\mathbb{R}^d$, i.e $\langle v_i, v_j \rangle = \delta_{ij}$, onde $\langle \cdot, \cdot \rangle$ é o produto interno usual em $\mathbb{R}^d$, $\delta_{ii} = 1$ e $\delta_{ij} = 0$ se $i \neq j$. Temos assim,

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2 \geq \max_j \|Av_j\|_2 = \max_j |\lambda_j| \, \|v_j\|_2 = \|A\|_2.$$

Seja $x = \sum_j x_j v_j$ com $\|x\|_2^2 = \sum_j |x_j|^2 = 1$. Então,

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle = \sum_{ij} \lambda_i \lambda_j x_i x_j \langle v_i, v_j \rangle = \sum_i \lambda_i^2 x_i^2 \leq \|A\|_2^2.$$

Concluindo, $\|A\|_2 \leq \|A\| \leq \|A\|_2$, ou seja $\|A\| = \|A\|_2 = \rho(A)$ para qualquer matriz simétrica $A$.

**Proposição 2.22.** *Para quaisquer normas de $E$ e $F$ temos que $\|A\| \geq \rho(A)$.*

*Demonstração.* Como para qualquer vector próprio $v$ de $A$ de norma 1 com valor próprio $\lambda$ temos $\|A\| \geq \|Av\| = |\lambda| \, \|v\| = |\lambda|$, temos que $\|A\|$ é maior ou igual ao maior valor próprio em valor absoluto (raio espectral). □

**Proposição 2.23.** *Se $A \in \mathcal{M}_{d\times d'}(\mathbb{R})$ e $B \in \mathcal{M}_{d'\times d''}(\mathbb{R})$, então*

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

*Demonstração.* Se $B = 0$ então $\|AB\| = \|B\| = 0$ e a desigualdade verifica-se. Supondo que $B \neq 0$, podemos garantir a existência de $x \neq 0$ tal que $Bx \neq 0$. Logo,

$$\begin{aligned} \|AB\| &= \sup_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \sup_{Bx\neq 0} \frac{\|ABx\| \, \|Bx\|}{\|Bx\| \, \|x\|} \\ &\leq \sup_{Bx\neq 0} \frac{\|ABx\|}{\|Bx\|} \sup_{Bx\neq 0} \frac{\|Bx\|}{\|x\|} \leq \|A\| \, \|B\|. \end{aligned}$$

□

### 2.3.3 Estimação de erros

Consideremos em vez da equação $A\,x = b$, a equação aproximada $\widetilde{A}\,\widetilde{x} = \widetilde{b}$. Queremos estimar o erro absoluto cometido $E = \|\widetilde{x} - x\|$ e o respectivo erro relativo $e = \|\widetilde{x} - x\|/\|x\|$.

Se considerarmos somente um erro inicial em $b$, i.e. $\widetilde{A} = A$, temos

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

Logo, os erros são dados por

$$\|\widetilde{x} - x\| \leq \|A^{-1}\|\,\|\widetilde{b} - b\| \quad \text{e} \quad \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \|A^{-1}\|\,\|A\|\frac{\|\widetilde{b} - b\|}{\|b\|},$$

onde usámos $\|b\| \leq \|A\|\,\|x\|$.

**Observação 2.24.** Note que o número de condição $cond(A) = \|A^{-1}\|\,\|A\|$ é determinante na estimação da grandeza do erro. Além disso, como $\|A^{-1}A\| = \|I\| = 1 \leq \|A\|\,\|A^{-1}\|$, temos sempre $cond(A) \geq 1$. Se $cond(A) \gg 1$ a solução da equação $A\,x = b$ é sensível às perturbações de $b$ (matriz $A$ mal condicionada). Se $cond(A) \simeq 1$, a solução é robusta (bem condicionada).

**Exemplo 2.25.**

1. Seja $A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$. Temos $\|A\|_2 = \|A^{-1}\|_2 = 1$.

2. Seja $A = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$. Agora $\|A\|_2 = \|A^{-1}\|_2 = 2$.

**Exercício 2.26.** *Resolva a equação $Ax = b$ com $A = \begin{bmatrix} 10 & 7 \\ 7 & 5 \end{bmatrix}$, $b = (32, 23)$ e $b = (32.1, 22.9)$. Calcule as duas soluções e conclua que a matriz $A$ é mal condicionada.*

Consideremos agora o caso de perturbações na matriz $A$. Neste caso estamos a assumir que $\widetilde{b} = b$. Logo, $\widetilde{A}\widetilde{x} = Ax$.

**Proposição 2.27.** *Seja $B \in \mathcal{M}_{d\times d}(\mathbb{R})$ tal que $\|B\| < 1$. Então, $I + B$ é invertível,*

$$(I + B)^{-1} = \sum_{n\geq 0}(-B)^n \quad e \quad \|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

*Demonstração.* Supondo que $I + B$ não é invertível, i.e. $\det(I + B) = 0$, temos que $\lambda = -1$ é valor próprio de $B$. Então, $Bv_1 = -v_1$ para o vector próprio $v_1$ respectivo. Finalmente, $\|B\| = \sup_x \|Bx\|/\|x\| \geq \|Bv_1\|/\|v_1\| = 1$.

Para verificar a fórmula da inversa basta calcular

$$(I+B)\sum_{n\geq 0}(-B)^n = \sum_{n\geq 0}(-B)^n - \sum_{n\geq 1}(-B)^n = I.$$

A estimativa da norma segue de

$$\|(I+B)^{-1}\| = \left\|\sum_{n\geq 0}(-B)^n\right\| \leq \sum_{n\geq 0}\|B\|^n = \frac{1}{1-\|B\|}$$

pois $\|B\| < 1$. □

Pela proposição acima aplicada a $\widetilde{A} = [I + (\widetilde{A} - A)A^{-1}]A$, se $\|\widetilde{A} - A\| < \|A^{-1}\|^{-1}$ a matriz $\widetilde{A}$ é invertível. Logo, a equação $\widetilde{A}\widetilde{x} = b$ tem solução e

$$\widetilde{x} - x = (\widetilde{A}^{-1}A - I)x$$

Ou seja, os erros são estimados por

$$\|\widetilde{x} - x\| \leq \|\widetilde{A}^{-1}\|\,\|\widetilde{A} - A\|\,\|x\| \quad \text{e} \quad \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \|\widetilde{A}^{-1}\|\,\|\widetilde{A} - A\|.$$

**Exercício 2.28.** *Considere o caso geral, i.e.* $\widetilde{A} \neq A$ *e* $\widetilde{b} \neq b$. *Determine estimativas para os erros associados à solução.*

# 3 Interpolação polinomial

Neste capítulo vamos estudar soluções para o seguinte problema. São-nos dados $n+1$ pontos $(x_i, y_i)$ em $\mathbb{R}^2$, $i = 0, \ldots, n$, correspondendo a pontos do gráfico de uma função desconhecida definida num intervalo $[a, b]$ [7]. Desta forma todas as abcissas $x_i$ são distintas. Encontrar a função original é tarefa impossível pois existem infinitas funções todas coincidindo nos $n+1$ pontos dados. Porém, podemos fazer algumas escolhas baseadas em critérios "razoáveis".

Aos valores $x_i$ chamamos nós interpoladores e a $y_i$ valores nodais. Dizemos que $f\colon [a, b] \to \mathbb{R}$ é uma função interpoladora se $f(x_i) = y_i$, $i = 0, \ldots, n$.

Os polinómios são bons candidatos a funções interpoladoras pois são definidos por um número finito de coeficientes reais (apropriado para o cálculo numérico) e aproximam relativamente bem funções contínuas (ver teorema de Weierstrass abaixo). Deste modo vamo-nos restringir ao caso das interpolações polinomais.

[7] Este é um problema típico nas ciências experimentais, quando se efectuam medições de uma determinada grandeza relativamente a uma variável. E.g. medições da pressão atmosférica a diferentes valores de altitude, o valor do PIB nacional em cada mês, a distância percorrida por uma bola de futebol variando a impulsão inicial, o gasto de combustível de um veículo variando o peso, etc.

## 3.1 Aproximação por polinómios

Seja $\mathcal{P}$ o conjunto de todos os polinómios em $\mathbb{R}$. Observe que $P \in \mathcal{P}$ é uma função analítica[8] em $\mathbb{R}$. Considere ainda a norma uniforme em $C^0([a,b])$ dada por

$$\|f\|_{C^0} = \sup_{x\in[a,b]} |f(x)|.$$

**Teorema 3.1** (Weierstrass)**.** *Para quaisquer $\varepsilon > 0$ e $f \in C^0([a,b])$, existe $P \in \mathcal{P}$ tal que $\|f - P\|_{C^0} < \varepsilon$.*

**Observação 3.2.** Note que o grau do polinómio $P$ depende de $\varepsilon$ e de $f$. Pode ser um valor muito elevado.

**Exercício 3.3.**

1. *Detenha-se por alguns momentos a apreciar o resultado do Teorema de Weierstrass.*
2. **Escreva uma demonstração deste resultado clássico (recorra à biblioteca).*
3. *Mostre que $f(x) = x + 10^{-5}\sin(10^{10}x)$ e $g(x) = x$ estão à distância $10^{-5}$ na norma uniforme.*
4. *Calcule a distância entre $f$ e $g$ para a norma $C^1$ dada por $\|f\|_{C^1} = \max\{\|f\|_{C^0}, \|f'\|_{C^0}\}$.*

**Exercício 3.4.**

1. *Confira que o número de operações algébricas de multiplicação e soma envolvidas no cálculo do valor de um polinómio com grau $n$ na forma $P(x) = \sum_{0\le j\le n} a_j x^j$, é dado por:*
   - *Número de multiplicações: $n(n+1)/2$*
   - *Número de somas: $n$*
   - *Número total: $n(n+3)/2$*
2. *Mostre que*

$$P(x) = a_0 + x\left(a_1 + x\left(a_2 + x\left(\cdots + xa_n\right)\right)\right),$$

   *e verifique que neste caso o número de operações reduz-se a:*
   - *Número de multiplicações: $n$*
   - *Número de somas: $n$*
   - *Número total: $2n$*

[8] Uma função $f$ diz-se analítica se $f \in C^\infty(\mathbb{R})$ e a sua série de Taylor converge (é igual a $f$) no seu domínio. A classe de funções analíticas representa-se por $C^\omega$.

## 3.2 Espaço de polinómios de grau $\leq n$

Considere o conjunto $\mathcal{P}_n$ dos polinómios de grau $\leq n$. Note que $\mathcal{P} = \bigcup_{n\geq 0} \mathcal{P}_n$.

**Exercício 3.5.** *Mostre que $\mathcal{P}_n$ é um espaço vectorial.*

Se considerarmos a base de $\mathcal{P}_n$ dada por

$$\{M_0, \ldots, M_n\} \quad \text{com} \quad M_j(x) = x^j,$$

qualquer polinómio $P \in \mathcal{P}_n$ é uma combinação linear na forma

$$P = \sum_{j=0}^{n} a_j M_j,$$

onde $a_j \in \mathbb{R}$. A dimensão de $\mathcal{P}_n$ é assim $n+1$. Outras bases de $\mathcal{P}_n$ serão úteis para o cálculo numérico e para o problema de interpolação polinomial, como iremos ver mais abaixo.

Dados $x_0, \ldots, x_n \in \mathbb{R}$ distintos, os polinómios de Lagrange são definidos por

$$L_j(x) = \prod_{k\neq j} \frac{x - x_k}{x_j - x_k}, \quad j = 0, \ldots, n.$$

Note que

$$L_j(x_i) = \prod_{k\neq j} \frac{x_i - x_k}{x_j - x_k} = \delta_{i,j}$$

e o grau de $L_j$ é $n$.

**Proposição 3.6.** *$\{L_0, \ldots, L_n\}$ é uma base de $\mathcal{P}_n$.*

*Demonstração.* Se $\sum_j c_j L_j = 0$, então no ponto $x_i$ temos $\sum_j c_j \delta_{i,j} = c_i = 0$ para qualquer $i = 0, \ldots, n$. Ou seja, os $n+1$ polinómios de Lagrange são linearmente independentes e geram $\mathcal{P}_n$. □

**Exercício 3.7.** *Escreva a matriz mudança de base entre as bases acima descritas para $n = 2$.*

Sejam $x_0, \ldots, x_n \in \mathbb{R}$ distintos. Outra base de $\mathcal{P}_n$ que utilizaremos mais à frente é dada pelos polinómios de Newton[9]:

$$N_j(x) = \prod_{k=0}^{j-1} (x - x_k), \quad j = 0, \ldots, n.$$

É fácil de verificar que

$$N_j(x_i) = 0 \quad \text{se} \quad i < j,$$

e que o grau de $N_j$ é $j$.

[9] Usamos a convenção $\prod_{k=0}^{-1} u_k = 1$.

**Proposição 3.8.** *$\{N_0, \ldots, N_n\}$ é uma base de $\mathcal{P}_n$.*

**Exercício 3.9.** *Prove-o.*

**Exercício 3.10.** *Escreva $P(x) = x^3 - x^2 + 2$ usando a base dos polinómios de Newton com $x_0 = 1$, $x_1 = -1$ e $x_2 = 0$.*

**Exemplo 3.11.** Queremos construir $P \in \mathcal{P}_1$ tal que

$$P(5000) = 0.1234 \quad \text{e} \quad P(5001) = -0.8766.$$

1. Solução exacta. Escrevendo $P(x) = a_0 + a_1 x$ e calculando-o nos pontos acima, obtemos o sistema

$$\begin{cases} a_0 + 5000 a_1 = 0.1234 \\ a_0 + 5001 a_1 = -0.8766 \end{cases}$$

   com solução $a_0 = 5000.1234$ e $a_1 = -1$.

2. Solução numérica usando representação em ponto flutuante com parâmetros $p = 4$ e $q = q' = 10$. A solução obtida é $\widetilde{a}_0 = 0.5 \times 10^4$ e $\widetilde{a}_1 = -1$. Porém, o polinómio assim determinado $\widetilde{P}(x) = \widetilde{a}_0 + \widetilde{a}_1 x$ calculado nos pontos indicados dá valores muito diferentes dos esperados:

$$\widetilde{P}(5000) = 0 \quad \text{e} \quad \widetilde{P}(5001) = -1.$$

3. Repetindo a alínea anterior usando a base dos polinómios de Newton $P(x) = b_0 + b_1(x - 5000)$, obtemos agora $\widetilde{P}(x) = 0.1234 - (x - 0.5 \times 10^4)$ correspondendo ao valor exacto.

## 3.3 Existência e unicidade do polinómio interpolador

Dizemos que $P$ é um polinómio interpolador se $P \in \mathcal{P}_n$ e $P(x_i) = y_i$ para qualquer $i$.

Escolhendo uma base $\{K_0, \ldots, K_n\}$ de $\mathcal{P}_n$ escrevemos $P \in \mathcal{P}_n$ na forma

$$P = \sum_{j=0}^{n} c_j K_j.$$

O polinómio interpolador é determinado pelas $n + 1$ equações $P(x_i) = y_i$. Podemos assim calcular os coeficientes $c_j$ resolvendo as equações lineares

$$\sum_j c_j K_j(x_i) = y_i, \qquad i = 0, \ldots, n.$$

Fazendo corresponder $a_{i,j} = K_j(x_i)$ aos coeficientes de uma matriz $A$ com dimensão $(n + 1) \times (n + 1)$, as equações anteriores reduzem-se a

$$Ac = y$$

onde $c = (c_0, \ldots, c_n)$ e $y = (y_0, \ldots, y_n)$.

A escolha da base de $\mathcal{P}_n$ pode levar a uma grande simplificação do sistema. De facto, lembrando que os polinómios de Lagrange verificam $L_j(x_i) = \delta_{i,j}$, neste caso a matriz $A$ é a identidade. Então, $c = y$. Está assim demonstrado o teorema seguinte.

**Teorema 3.12** (Fórmula de Lagrange)**.** *O polinómio interpolador existe, é único e é dado por*

$$P = \sum_{j=0}^{n} y_j L_j.$$

**Exercício 3.13.** *Calcule a solução de $Ac = y$ usando a base $\{M_0, \ldots, M_n\}$ de $\mathcal{P}_n$. A matriz $A$ neste caso chama-se matriz de Vandermonde.*

**Exercício 3.14.** *Mostre que se o grau do polinómio interpolador fosse superior a $n$, então não seria único.*

**Exercício 3.15.** *Calcule o número de multiplicações, divisões e somas para determinar o valor num dado ponto do polinómio interpolador usando a fórmula de Lagrange. Mostre que o total de operações é da ordem de $n^2$.*

**Exemplo 3.16.** Queremos encontrar o polinómio interpolador $P$ para os seguintes dados:

| $x$ | 0 | 1 | 3 | 4 |
|---|---|---|---|---|
| $y$ | 1 | -1 | 1 | 2 |

Então,

$$\begin{aligned} L_0(x) &= -\frac{1}{12}(x-1)(x-3)(x-4) \\ L_1(x) &= \frac{1}{6}x(x-3)(x-4) \\ L_2(x) &= -\frac{1}{6}x(x-1)(x-4) \\ L_3(x) &= \frac{1}{12}x(x-1)(x-3). \end{aligned}$$

Finalmente, $P = L_0 - L_1 + L_2 + 2L_3$.

**Exercício 3.17.** *Considere a função $\Gamma\colon \mathbb{R}^+ \to \mathbb{R}$ dada por*

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt.$$

1. *Mostre que $\Gamma(n) = (n-1)!$, $n \in \mathbb{N}$.*
2. *Calcule valores aproximados para $\Gamma(3/2)$ usando a fórmula de Lagrange.*

**Observação 3.18.** Repare que quando acrescentamos mais um ponto aos nossos dados, temos que recalcular todos os polinómios de Lagrange. Isto porque cada $L_i$ depende de todos os nós simultaneamente. Esta desvantagem é removida usando a fórmula de Newton (ver secção seguinte).

## 3.4 Fórmula de Newton

Considere agora a base $\{N_0, \ldots, N_n\}$ de $\mathcal{P}_n$ definida pelos polinómios de Newton. Agora, a matriz $A$ é triangular inferior pois

$$a_{i,j} = \begin{cases} 0, & i < j \\ N_j(x_i), & \text{c.c.} \end{cases}$$

A solução de $Ac = y$ é assim

$$c_j = \frac{y_j - \sum_{k=0}^{j-1} c_k a_{i,k}}{a_{j,j}}.$$

Fica assim demonstrado o teorema seguinte.

**Teorema 3.19** (Fórmula de Newton)**.** *O polinómio interpolador é*

$$P = \sum_{j=0}^{n} c_j N_j,$$

*onde*

$$c_j = \frac{y_j - \sum_{k=0}^{j-1} c_k N_k(x_i)}{N_j(x_j)}.$$

**Observação 3.20.** Os coeficientes $c_j$ são chamados **diferenças divididas**. Note que $c_j$ apenas depende de $x_0, \ldots, x_j$.

**Proposição 3.21.** *Sejam $(x_i, y_i)$, $i = 0, \ldots, n+1$, com todos os $x_i$ distintos. Se $P_n \in \mathcal{P}_n$ é o polinómio interpolador de $(x_i, y_i)$, $i = 0, \ldots, n$, então o polinómio interpolador $P_{n+1} \in \mathcal{P}_{n+1}$ de $(x_i, y_i)$, $i = 0, \ldots, n+1$ é*

$$P_{n+1} = P_n + c_{n+1} N_{n+1}.$$

**Exercício 3.22.** *Prove-o.*

**Observação 3.23.** Quando acrescentamos mais nós interpoladores (de forma a controlarmos a aproximação polinomial, ou porque obtivemos mais dados experimentais), não temos que recalcular todos os coeficientes.

## 3.5 Cálculo de diferenças divididas

Do que vimos acima temos que as diferenças divididas são dadas por

$$c_j = \frac{y_j - P_{j-1}(x_i)}{N_j(x_j)}.$$

Queremos obter uma fórmula iterativa para o cálculo destes coeficientes.

Seja

$$[y_i] = y_i$$

para qualquer $i = 0, \ldots, n$. Dados $0 \leq k < k' \leq n$ definimos também

$$[y_k, \ldots, y_{k'}] = \frac{[y_{k+1}, \ldots, y_{k'}] - [y_k, \ldots, y_{k'-1}]}{x_{k'} - x_k}.$$

**Teorema 3.24.** $c_j = [y_0, \ldots, y_j]$.

**Exercício 3.25.** **Demonstre o teorema anterior.*

**Exemplo 3.26.** Na tabela seguinte apresentamos 4 nós e as correspondentes diferenças divididas.

| $x_i$ | $y_i = [y_i]$ | $[y_i, y_{i+1}]$ | $[y_i, y_{i+1}, y_{i+2}]$ | $[y_i, y_{i+1}, y_{i+2}, y_{i+3}]$ |
|---|---|---|---|---|
| 1 | 1 | | | |
| | | 0 | | |
| 2 | 1 | | $\frac{1}{2}$ | |
| | | 1 | | $\frac{1}{3}$ |
| 3 | 2 | | $\frac{3}{2}$ | |
| | | 4 | | |
| 4 | 6 | | | |

Assim, $P_3(x) = 1 + 0(x-1) + \frac{1}{2}(x-1)(x-2) + \frac{1}{3}(x-1)(x-2)(x-3)$. A partir dos valores na tabela, podemos considerar várias combinações de pontos sem recorrer a mais cálculos. Por exemplo:

1. Usando $x_1 = 2$ e $x_2 = 3$ calculamos $P_1(x) = 1 + (x-2)$.
2. Usando $x_1 = 2$, $x_2 = 3$ e $x_3 = 4$ temos $P_2(x) = P_1(x) + \frac{3}{2}(x-2)(x-3)$.

**Exercício 3.27.** *Escreva um código para calcular um polinómio interpolador. Use-o para calcular o polinómio interpolador para as cotações (no fecho da sessão diária) das acções da Apple Inc. (ou outra empresa ou índice à sua escolha) durante os últimos 30 dias*[10]*. Desenhe o gráfico do polinómio interpolador extrapolando para os próximos 30 dias.*

**Exercício 3.28.** *Para a função do Exercício* 3.17*, obtenha valores aproximados para* $\Gamma(3/2)$ *usando diferenças divididas.*

## 3.6 Erro de interpolação polinomial

Se estivermos a estudar uma função $f$, o erro de interpolação polinomial é dado naturalmente por

$$E(x) = f(x) - P(x).$$

onde $P$ é o polinómio interpolador. A função é desconhecida, pelo que a estimativa do erro terá que ser baseada nalguma hipótese sobre $f$. Assumir que $f \in C^{n+1}$ é por vezes razoável. Iremos considerar este caso e assim estimar o erro de interpolação.

**Observação 3.29.**

[10] Consulte https://finance.yahoo.com/quote/AAPL/history?p=AAPL para obter os dados.

1. $E(x_i) = 0$ para qualquer $i = 0, \ldots, n$.

2. Se $f \in \mathcal{P}_n$ então $f = P$ pela unicidade do polinómio interpolador e $E = 0$.

**Teorema 3.30.** *Seja $f \in C^{n+1}([a,b])$ e $P$ o polinómio interpolador. Então, para qualquer $x \in [a,b]$ existe $\xi \in ]a,b[$ tal que*

$$E(x) = \frac{N_{n+1}(x)}{(n+1)!} f^{(n+1)}(\xi).$$

*Demonstração.* Note que $N_{n+1}(x_i) = 0$ e $E(x_i) = 0$ para qualquer $i = 0, \ldots, n$. Assuma agora que $x$ não é um nó interpolador. Assim $N_{n+1}(x) \neq 0$. Seja

$$F(t) = E(t) - cN_{n+1}(t) \quad \text{com} \quad c = \frac{E(x)}{N_{n+1}(x)}.$$

Então, $F(x_i) = 0$ e também $F(x) = 0$. Ou seja, $F$ é uma função $C^{n+1}$ com pelo menos $n+2$ zeros. Logo, $F^{(n+1)}$ tem pelo menos um zero $\xi$. Isto é,

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P^{(n+1)}(\xi) - cN^{(n+1)}(\xi) = 0.$$

Como $P^{(n+1)} = 0$ uma vez que o grau de $P$ é $\leq n$, e $N^{(n+1)} = (n+1)!$, obtemos

$$c = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Usando a definição de $c$ completamos a demonstração. □

Podemos então estimar o erro da seguinte forma:

$$\|E\|_{C^0} \leq \frac{\|N_{n+1}\|_{C^0}}{(n+1)!} \|f^{(n+1)}\|_{C^0}.$$

**Exercício 3.31.** *Mostre que se $|x_{i+1} - x_i| \leq h$, então $\|N_{n+1}\|_{C^0} \leq n! h^{n+1}$. Aproveite para estimar o erro de interpolação, obtendo*

$$\|E\|_{C^0} \leq \frac{h^{n+1}}{n+1} \|f^{(n+1)}\|_{C^0}.$$

*Observe que se $f \in C^\infty$ e $\|f^{(n)}\|_{C^0} \leq C$ para qualquer $n$ e $h \leq 1$, então o erro de interpolação converge para 0 quando se considera o número de nós arbitrariamente grande.*

**Exemplo 3.32.** Seja $f(x) = \sin(x)$ no intervalo $[0, 2\pi]$. Neste caso $\|f^{(n)}\|_{C^0} = 1$ para qualquer $n \in \mathbb{N}$. Assim, desde que se considerem nós equidistantes, a uma distância entre consecutivos não superior a 1, conseguimos obter aproximações polinomiais de $f$ tão boas quanto se queiram bastando para isso ir acrescentando nós.

## 3.7 Nós de Chebyshev

Se numa determinada aplicação for possível escolher os nós onde interpolamos a função[11], então podemos controlar de melhor forma o factor $\|N_{n+1}\|_{C^0}/(n+1)!$ no erro. Uma abordagem é através dos nós de Chebyshev, que definimos abaixo.

Os polinómios de Chebyshev $T_n \in \mathcal{P}_n$, $n \geq 0$, são definidos por recorrência da seguinte forma

$$T_0(x) = 1, \qquad T_1(x) = x \quad \text{e} \quad T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x), \quad x \in \mathbb{R}.$$

Iremos mostrar abaixo que os seus zeros correspondem à melhor escolha de nós de interpolação.

**Proposição 3.33.**

1. $T_n(-x) = (-1)^n T_n(x)$, $x \in \mathbb{R}$.

2. $$T_n(x) = \begin{cases} \cos(n \arccos x), & -1 \leq x \leq 1 \\ \cosh(n \operatorname{arcosh} x), & x \geq 1 \\ (-1)^n \cosh(n \operatorname{arcosh}(-x)), & x \leq -1. \end{cases}$$

3. *Para $n \geq 1$ os zeros de $T_n$ são dados por*

$$z_i = \cos\left[\frac{(2i+1)\pi}{2n}\right], \quad i = 0, \ldots, n-1.$$

4. *Em $[-1, 1]$ os extremantes de $T_n$ com $n \geq 1$ são dados por*

$$z'_k = \cos\left(\frac{k\pi}{n}\right), \quad k = 0, \ldots, n,$$

*com valores $T_n(z'_k) = (-1)^k$.*

**Exercício 3.34.** *Prove a proposição acima.*

Observe que todos os zeros de $T_n$ estão em $[-1, 1]$. São chamados nós de Chebyshev de ordem $n$.

**Observação 3.35.**

1. O termo de ordem $n$ de $T_n$ tem coeficiente $2^{n-1}$. Logo,

$$T_n(x) = 2^{n-1} \prod_{i=0}^{n-1} (x - z_i), \quad x \in \mathbb{R}.$$

[11] Se forem dados experimentais, basta colher as amostras apenas para determinados valores dos parâmetros.

2. $\sup_{x\in[-1,1]} |T_n(x)| = 1$.

Usando a transformação afim $h\colon [-1,1] \to [a,b]$ dada por

$$h(x) = \frac{b-a}{2}x + \frac{b+a}{2},$$

definimos os nós de Chebyshev de grau $n+1$ em $[a,b]$ por

$$\xi_i = h(z_i) = \frac{b-a}{2}\cos\left[\frac{(2i+1)\pi}{2(n+1)}\right] + \frac{b+a}{2}, \quad i = 0,\ldots,n,$$

correspondendo aos zeros do polinómio $T_n \circ h^{-1}(x)$.

**Proposição 3.36.** *Considere os nós de interpolação dados pelos nós de Chebyshev $\xi_0,\ldots,\xi_n$ de grau $n+1$ em $[a,b]$. Então*

$$\|N_{n+1}\|_{C^0} = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}.$$

*Demonstração.* Basta verificar que para qualquer $x \in [a,b]$ temos

$$N_{n+1}(x) = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1} T_{n+1} \circ h^{-1}(x)$$

e que $h^{-1}(x) \in [-1,1]$. □

**Observação 3.37.** O erro de interpolação usando os nós de Chebyshev pode ser assim estimado por

$$\|E\|_{C^0} \leq \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1} \frac{\|f^{(n+1)}\|_{C^0}}{(n+1)!}.$$

Resta mostar que $\|N_{n+1}\|_{C^0}$ toma o menor valor possível quando é construída usando os nós de Chebyshev.

**Proposição 3.38.** *Considere nós de interpolação $x_0,\ldots,x_n$ em $[a,b]$. Então*

$$\|N_{n+1}\|_{C^0} \geq \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}.$$

*Demonstração.* Para demonstrar esta desigualdade suponha que

$$\|N_{n+1}\|_{C^0} < \lambda_n. \tag{3.1}$$

com

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

O polinómio

$$Q = \lambda_n T_{n+1} \circ h^{-1} - N_{n+1}$$

tem grau $\leq n$ pois os termos de ordem $n+1$ cancelam-se. Então, usando os extremantes $z'_k$ de $T_{n+1}$ em $[-1,1]$ obtemos

$$Q(h(z'_k)) = (-1)^k \lambda_n - N_{n+1}(h(z'_k)), \quad k = 0, \ldots, n+1.$$

Devido a (3.1) $Q$ tem $n+1$ zeros. Porém, $Q$ só pode ser um polinómio de grau $n$ e simultaneamente ter $n+1$ zeros se $Q = 0$. Isto implica que $N_{n+1} = \lambda_n T_{n+1} \circ h^{-1}$ cuja norma contradiz (3.1). □

**Exercício 3.39.** *Calcule o polinómio interpolador da função* $f(x) = (x^2 + 1)^{-1}$ *em* $[-5, 5]$ *usando os nós de interpolação:*

1. $x_i = -5 + i$, $i = 0, \ldots, 10$.
2. *Os nós de Chebyshev em* $[-5, 5]$ *de grau 11.*

### 3.8 Splines cúbicos

Um método muito popular de interpolação baseia-se no uso de splines cúbicos (ou somente splines). Consiste em aproximar a função entre cada dois nós por um polinómio de grau $\leq 3$ de forma que a função global seja $C^2$. Ou seja, uma spline é uma função $S(x) = S_i(x)$ para $x \in [x_i, x_{i+1}[$, $i = 1, \ldots, n$, onde $S_i$ são polinómios de grau $\leq 3$ tais que $S \in C^2$ e $S(x_i) = y_i$.

A definição de $S$ é dada então por:

$$\begin{aligned} S_i(x) = & \frac{M_{i-1}(x_i - x)^3 - M_i(x_{i-1} - x)^3}{6(x_i - x_{i-1})} \\ & + \frac{[6y_{i-1} - M_{i-1}(x_i - x_{i-1})^2](x_i - x) - [6y_i - M_i(x_i - x_{i-1})^2](x_{i-1} - x)}{6(x_i - x_{i-1})}. \end{aligned}$$

Para determinar os coeficientes $M_0, \ldots, M_n$ é necessário impôr as $n-1$ condições: $S'_i(x_i^-) = S'_{i+1}(x_{i+1}^+)$. Faltam mais duas condições para podermos determinar todos os $M_i$. Para isso temos que usar condições fronteira. Por exemplo, fixando $S'_1(x_0)$ e $S'_n(X_n)$ ou então $S''_1(x_0)$ e $S''_n(x_n)$.

## 4 Métodos numéricos para equações não lineares

Após termos tratado a resolução de equações lineares, queremos agora desenvolver métodos para obter aproximações de soluções de equações não lineares, nomeadamente encontrar zeros de funções. Seja $f\colon D \to \mathbb{R}^d$ com $D \subset \mathbb{R}^d$. O nosso problema é então resolver a equação

$$f(z) = 0,$$

onde $z \in D$ é zero de $f$. Esta tarefa não é trivial. Por exemplo, para $f(x) = e^x + x$ é simples verificar que existe um único zero visto que é uma função contínua com $f(-1) < 0 < f(0)$ e crescente ($f'(x) > 0$). Porém, a determinação do zero não é simples.

Se uma função não for contínua, pelo menos no intervalo que estejamos interessados em estudar, então torna-se impossível determinar os zeros. Assumimos assim que as funções estudadas são contínuas nos respectivos domínios.

Cada um dos métodos seguintes consiste num algoritmo que iterado produz uma sucessão de pontos $x_n$ que aproximam-se assimptoticamente do zero $z$ de $f$. O erro de aproximação a cada passo $n$ é dado por

$$e_n = z - x_n.$$

A forma de determinarmos quando um algoritmo deve parar baseia-se no teste a cada passo do valor do erro. Isto é, dado $\varepsilon > 0$ queremos determinar o tempo de paragem $N$ tal que para $n \geq N$ obtemos $|e_n| \leq \varepsilon$.

No caso da função ser desconhecida, o método mais popular para decidir quando parar a iteração baseia-se no teste da seguinte proposição ao $n$-ésimo passo:

$$|x_n - x_{n-1}| < \tau'|x_n| + \tau,$$

para uma escolha de constantes $\tau, \tau'$. Por exemplo, estas constantes podem ser definidas a partir da condição $fl(x_n) = fl(x_{n-1})$.

## 4.1 Método da bissecção

Este método é restrito ao caso $d = 1$ pois baseia-se num resultado de análise em $\mathbb{R}$: o teorema do valor intermédio (teorema de Bolzano).

Considere $f \in C^0([a, b])$ tal que $f(a)\, f(b) < 0$, i.e. os sinais de $f$ nos extremos do intervalo são contrários. Pelo teorema do valor intermédio existe pelo menos um zero em $]a, b[$. Este facto leva-nos a considerar um algoritmo simples de pesquisa de zeros. Denotando $I_0 = [a_0, b_0]$ com $a_0 = a$ e $b_0 = b$, testamos o sinal de $f$ no valor médio de $I_0$ dado por $\frac{a_0+b_0}{2}$. Escolhemos então um novo intervalo $I_1 = [a_1, b_1]$ com metade da largura do anterior e onde temos seguramente um zero (por ter extremos com sinais contrários). Iterando este procedimento obtemos uma sucessão de intervalos

$$I_n = [a_n, b_n] \subset I_{n-1}, \qquad n \in \mathbb{N},$$

contendo um zero de $f$, onde o par dos extremos de cada intervalo é dado por

$$(a_n, b_n) = \begin{cases} (a_{n-1}, \frac{a_{n-1}+b_{n-1}}{2}), & f(a_{n-1})\, f(\frac{a_{n-1}+b_{n-1}}{2}) < 0 \\ (\frac{a_{n-1}+b_{n-1}}{2}, b_{n-1}), & f(a_{n-1})\, f(\frac{a_{n-1}+b_{n-1}}{2}) \geq 0. \end{cases}$$

A aproximação a $z$ em cada passo é escolhida como o ponto médio de $I_n$:

$$x_n = \frac{a_n + b_n}{2}.$$

Note que se obtivermos $f(a_n) = 0$ ou $f(b_n) = 0$ para algum $n$, então o zero está determinado e a iteração pode terminar.

**Exercício 4.1.** *Implemente computacionalmente o método da bissecção para determinar zeros de uma função $f \in C^0([a,b])$. Use-o para calcular um zero de $f(x) = e^x + x + 10$.*

O próximo teorema determina a convergência deste algoritmo e o erro de aproximação a cada passo $n$.

**Teorema 4.2.** *Seja $f \in C^0([a,b])$ tal que $f(a)\,f(b) < 0$. Então, $\lim x_n$ é um zero de $f$ e*

$$|e_n| \leq \frac{b-a}{2^{n+1}}.$$

*Demonstração.* Note que $a_n$ e $b_n$ são sucessões limitadas e monótonas, logo convergentes. Assim $\frac{a_n+b_n}{2}$ também converge sendo o limite denominado $z$. Por outro lado, como a largura de $I_n$ é dada por $b_n - a_n \leq \frac{1}{2}(b_{n-1} - a_{n-1}) \leq \cdots \leq \frac{1}{2^n}(b-a)$, os limites de $a_n$ e $b_n$ são também iguais a $z$. O erro é estimado por metade da largura de $I_n$, i.e. $|e_n| \leq \frac{1}{2}(b_n - a_n)$.

Falta verificar que $f(z) = 0$. Ora, como $f(a_n)\,f(b_n) < 0$, usando o facto que $f$ é contínua, $\lim f(a_n)\,f(b_n) = f(\lim a_n)\,f(\lim b_n) = f(z)^2 \leq 0$. Ou seja, $f(z) = 0$. □

**Exercício 4.3.** *Mostre que o tempo de paragem usando o método da bissecção é dado por $N = \log(|e_0|/\varepsilon)/\log 2$.*

## 4.2 Método de Newton

O método de Newton corresponde a uma iteração onde em cada passo determinamos o zero da linearização de $f$. Este passo é trivial pois sabemos calcular zeros de sistemas lineares desde que não seja um sistema singular. Como iremos ver, o método de Newton distingue-se pela sua rapidez de convergência para um zero de $f$. O preço a pagar é a necessidade de obter uma boa condição inicial $x_0$. Ou seja, a convergência só é boa se $x_0$ estiver já numa vizinhança suficientemente pequena de $z$.

Recorde que a fórmula de Taylor de 1ª ordem em $x_0$ para uma função $C^2$ é dada por

$$f(x) = f(x_0) + Df(x_0)\,(x - x_0) + R_2(x, x_0),$$

onde a matriz derivada é

$$Df(x_0) = \left[\frac{\partial f_i}{\partial x_j}(x_0)\right]_{i,j=1,\ldots,d}$$

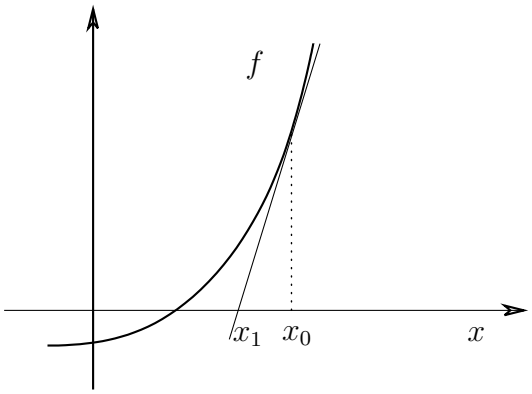


Figura 3: Exemplo do primeiro passo do método de Newton em dimensão 1.

e $R_2$ é o resto de ordem 2, i.e. $\|R_2(x, x_0)\| \leq C\|x - x_0\|^2$ para algum $C$ dependendo de $x_0$.

Queremos resolver $f(z) = 0$, mas em vez disso calculamos o zero $x_1$ da linearização de $f$:

$$f(x_0) + Df(x_0)\,(x_1 - x_0) = 0$$

dado por

$$x_1 = N(x_0) = x_0 - Df(x_0)^{-1} f(x_0)$$

desde que a derivada de $f$ seja invertível em $x_0$. Observe agora que $f(x_1) = R_2(x_1, x_0)$ tem norma da ordem $\|x_1 - x_0\|^2$.

**Observação 4.4.** O caso $d = 1$ não se distingue do caso $d \geq 2$. Porém, em dimensão 1 temos também uma interpretação geométrica do algoritmo, tendo em conta que a linearização de $f$ é a recta tangente ao gráfico da função em cada ponto. Ou seja, $x_1$ é o ponto de intersecção entre a recta e o eixo horizontal (ver Figura 3).

Iterando o procedimento anterior obtemos o algoritmo:

$$x_{n+1} = N(x_n), \qquad n \geq 0,$$

para uma escolha inicial $x_0$. Como $N$ é contínua, se $\lim x_n$ existe este é necessariamente um zero de $f$. De facto, $\lim N(x_n) = N(\lim x_n) = \lim x_n - Df(\lim x_n)^{-1} f(\lim x_n)$. Assumindo que a derivada é invertível nesse ponto, a igualdade $\lim x_{n+1} = \lim N(x_n)$ verifica-se sse $f(\lim x_n) = 0$.

Para determinarmos a convergência e uma estimativa do erro, consideremos primeiro o caso unidimensional.

**Teorema 4.5.** *Seja $f \in C^2([a, b])$ tal que $0 < m \leq |f'(x)| \leq M$ e $|f''(x)| \leq M'$, $x \in ]a, b[$, e $c = M'(2m)^{-1}$. Para $x_0$ na vizinhança de raio $c^{-1}$ de um zero $z$ de $f$ temos que*

1. $|e_{n+1}| \leq c\,|e_n|^2$,

2. $|e_n| \leq c^{-1}(c\,|e_0|)^{2^n}$,

3. $\lim x_n = z$.

*Demonstração.* Escrevendo a fórmula de Taylor de 1ª ordem em $x_n$ calculada no ponto $z$

$$f(z) = f(x_n) + f'(x_n)(z - x_n) + \frac{1}{2}f''(\xi)(z - x_n)^2 = 0$$

com $\xi$ entre $z$ e $x_n$, obtemos

$$z = N(x_n) - \frac{f''(\xi)}{2f'(x_n)}(z - x_n)^2.$$

Logo,

$$|e_{n+1}| = |z - x_{n+1}| = \left|\frac{f''(\xi)}{2f'(x_n)}\right| |e_n|^2 \leq c|e_n|^2.$$

Por indução, temos que $|e_n| \leq c^{-1}(c|e_o|)^{2^n}$. Se $|e_o| < c^{-1}$, então $|e_n| \to 0$ e $\lim x_n = z$. □

**Teorema 4.6.** *Sejam $f \in C^1(D)$, $D \subset \mathbb{R}^d$ aberto e convexo, $z \in D$ zero de $f$, $Df(z)$ invertível, $\|Df(z)^{-1}\| \leq \beta$ e $\|Df(x) - Df(y)\| \leq \gamma\|x - y\|$, $x, y \in D$. Existe $c > 0$ tal que para $\|z - x_0\| < c^{-1}$,*

1. $\|e_{n+1}\| \leq c\|e_n\|^2$,

2. $\|e_n\| \leq c^{-1}(c\|e_0\|)^{2^n}$,

3. $\lim x_n = z$.

*Demonstração.* Como temos para um $y$ no segmento entre $z$ e $x_n$ (sendo $D$ convexo está garantido que $y \in D$) a fórmula de Taylor 1ª ordem dada por

$$0 = f(z) = f(x_n) + [Df(y) - Df(x_n) + Df(x_n)](z - x_n),$$

então

$$z - x_{n+1} = Df(x_n)^{-1}[Df(y) - Df(x_n)](z - x_n).$$

Acima assumimos que $x_n$ está suficientemente próximo de $z$ para que $Df(x_n)$ seja também invertível com norma majorada por $2\beta$. As estimativas seguem assim como no caso anterior do método de Newton para dimensão 1. □

**Exercício 4.7.** *Na demonstração anterior usámos o facto de uma matriz quadrada suficientemente próxima de outra invertível ainda é invertível. Demonstre-o usando a relação*

$$A^{-1} = (A - B + B)^{-1} = B^{-1}[(A - B)B^{-1} + I]^{-1}$$

*e $(I + C)^{-1} = \sum_{n\geq 0} C^n$ desde que $\|C\| < 1$.*

**Exercício 4.8.** *Compute numericamente zeros para a função* $f(x,y) = (1 + e^x \sin y - x, x^2 - e^{\sin y})$.

**Exercício 4.9.** *Calcule o tempo de paragem do método de Newton.*

Considere agora a seguinte modificação ao método de Newton. Seja $f \in C^1(D)$, $D \subset \mathbb{R}^d$ aberto e convexo, e $y \in D$ tal que $Df(y)$ é invertível. O método de Newton modificado é dado pela iteração da função

$$\widetilde{N}(x) = x - Df(y)^{-1} f(x).$$

Definimos assim a sucessão $x_n = \widetilde{N}(x_{n-1})$ para um ponto inicial $x_0$. A vantagem deste método é não termos que calcular a derivada (e invertê-a) em cada $x_n$. Basta escolher um ponto inicial.

**Exercício 4.10.** *Mostre a seguinte proposição. Se*

$$c = \|Df(y)^{-1}\| \sup_{x \neq x'} \frac{\|Df(x) - Df(x')\|}{\|x - x'\|} < \infty,$$

*então para qualquer* $x_0 \in D$ *temos que*

1. $\|z - \widetilde{N}(x_0)\| \leq c\|z - x_0\|^2$.
2. *Se* $x_n \in D$ *para qualquer* $n \in \mathbb{N}$, *então*

$$\|z - x_{n+1}\| \leq c^{-1}(c\|z - x_n\|)^{2^n}.$$

3. *Se* $x_0$ *está suficientemente perto de* $z$ *então* $\lim x_n = z$.

## 4.3 Método da secante

A interpretação geométrica do método de Newton para $d = 1$ adapta-se facilmente a um novo método onde não seja necessário utilizar a derivada de $f$ (no caso de ser desconhecida ou difícil de calcular). Assim, em vez de considerar a recta tangente ao gráfico, consideramos dois pontos no gráfico de $f$ com abcissas $x_0$ e $x_1$ e a recta que use esses dois pontos dada por

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

Esta recta é secante ao gráfico de $f$. A intersecção com o eixo horizontal é no ponto

$$x_2 = x_1 - f(x_1)\frac{x_1 - x_0}{f(x_1) - f(x_0)}.$$

É óbvio que se a secante for horizontal ($f(x_1) - f(x_0) = 0$) o método não funciona.

Iterando o procedimento acima descrito, obtemos uma sequência de aproximações dada por

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \qquad n \in \mathbb{N},$$

para as condições iniciais $x_0$ e $x_1$. Como no método de Newton, basta que $\lim x_n$ exista para que seja um zero de $f$.

**Teorema 4.11.** *Seja $f \in C^2([a,b])$ tal que $0 < m \leq |f'(x)| \leq M$ e $|f''(x)| \leq M'$, $x \in ]a,b[$, e $c = M(2m)^{-1}$. Então, para $x_0$ e $x_1$ suficientemente próximos de um zero $z$ de $f$, $\lim x_n = z$, $|e_{n+1}| \leq c|e_n|\,|e_{n-1}|$ e*

$$|e_n| \leq c^{-1}(c \max\{|e_0|, |e_1|\})^{F_n},$$

*onde $F_n$ é a sucessão de Fibonacci*[12] *dada por $F_{n+2} = F_{n+1} + F_n$ com $F_0 = F_1 = 1$.*

**Exercício 4.12.** **Demonstre o teorema anterior.*

## 4.4 Comparação de métodos

A ordem de convergência assimptótica de um algoritmo é definida como

$$p = \lim |\log |e_n||^{1/n}.$$

É simples verificar que no caso do método de Newton temos

$$p_{Newton} = 2,$$

e para o método da secante obtemos

$$p_{secante} = \frac{1+\sqrt{5}}{2} < 2.$$

Finalmente, o método da bissecção dá-nos

$$p_{bissec} = 1.$$

Esta grandeza mede a rapidez de convergência dos algoritmos. O método de Newton é claramente o mais rápido.

**Exemplo 4.13.** Queremos obter uma aproximação de $\pi$. Podemos usar a função $f(x) = \tan x$ e calcular um zero próximo de $x_0 = 3$ com uma precisão de 5 casas decimais; para o método da secante considerar também $x_1 = 4$ e para o método da bisecção consideramos o intervalo inicial $[3,4]$:

[12] Mostre por indução que $F_n = \frac{\sqrt{5}}{5}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right]$.

| $i$ | Newton | Secante | Bissecção |
|---|---|---|---|
| 0 | 3 | 3 | 3.5 |
| 1 | 3.14255 | 4 | 3.25 |
| 2 | 3.14159 | 3.15716 | 3.125 |
| 3 | 3.14159 | 3.13946 | 3.1875 |
| 4 | 3.14159 | 3.14159 | 3.15625 |
| 5 | 3.14159 | 3.14159 | 3.14063 |
| 6 | 3.14159 | 3.14159 | 3.14844 |
| 7 | 3.14159 | 3.14159 | 3.14456 |
| 8 | 3.14159 | 3.14159 | 3.14258 |
| 9 | 3.14159 | 3.14159 | 3.14161 |
| 10 | 3.14159 | 3.14159 | 3.14112 |
| 11 | 3.14159 | 3.14159 | 3.14137 |
| 12 | 3.14159 | 3.14159 | 3.14149 |
| 13 | 3.14159 | 3.14159 | 3.14155 |
| 14 | 3.14159 | 3.14159 | 3.14158 |
| 15 | 3.14159 | 3.14159 | 3.14160 |
| 16 | 3.14159 | 3.14159 | 3.14159 |

## 4.5 Problemas de ponto fixo

O problema de determinar um zero de uma função é equivalente ao de determinar um ponto fixo:

$$g(z) = z$$

Note que se escrevermos por exemplo $f(x) = g(x) - x$, o ponto fixo de $g$ é o zero de $f$. Outro exemplo é o método de Newton com $N(x) = x - Df(x)^{-1}f(x)$ tem um ponto fixo $z$ sse $f(z) = 0$ desde que a derivada seja invertível.

Pretendemos aqui encontrar condições suficientes para que a solução seja obtida simplesmente por iterações[13] consecutivas da função: $\lim g^n(x_0) = z$. De facto, esta é a ideia por detrás do método de Newton.

Uma função $g\colon D \to \mathbb{R}^d$ é *contractiva* em $D \subset \mathbb{R}^d$ sse existe $0 < \theta < 1$ tal que para qualquer $x_1, x_2 \in D$ temos

$$\|g(x_1) - g(x_2)\| \leq \theta \|x_1 - x_2\|.$$

**Observação 4.14.** Uma função unidimensional $g \in C^1([a,b])$ é contractiva sse

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

Para demonstrar esta proposição basta usar o teorema do valor médio.

[13] Recorde que $g^n = g \circ g \circ \cdots \circ g$ é a composição de $g$ $n$ vezes, e não o produto.

**Teorema 4.15** (do ponto fixo)**.** *Seja $g \in C^0(D, \mathbb{R}^d)$ contractiva num compacto $D \subset \mathbb{R}^d$ tal que $g(D) \subset D$. Então existe um único ponto fixo $z \in D$ e $\lim g^n(x_0) = z$ para qualquer $x_0 \in D$.*

*Demonstração.* Para $x_0 \in D$ temos que $x_n = g^n(x_0)$ está em $D$ para qualquer $n \in \mathbb{N}$. Temos então,

$$\|x_{n+1} - x_n\| = \|g(x_n) - g(x_{n-1})\| \leq \theta \|x_n - x_{n-1}\| \leq \cdots \leq \theta^n \|x_1 - x_0\|.$$

Logo $x_n$ é convergente e denotamos $z = \lim x_n$. Como $D$ é compacto, então $z \in D$. Por outro lado, como $g$ é contínua, $\lim g(x_n) = g(\lim x_n) = \lim x_{n+1}$, ou seja $g(z) = z$.

Falta provar que o ponto fixo $z$ é único. Supondo que existem $z_1 \neq z_2$ tais que $g(z_i) = z_i$. Então $\|z_1 - z_2\| = \|g(z_1) - g(z_2)\| < \|z_1 - z_2\|$, o que leva a uma contradição. □

**Exercício 4.16.** *Mostre que o erro de aproximação é dado por $\|x_n - z\| \leq \theta^n \|x_0 - z\|$.*

**Exercício 4.17.** *Dê exemplos de funções contractivas para as quais as conclusões do teorema do ponto fixo não são válidas.*

**Exemplo 4.18.** Queremos calcular $\sqrt{k}$ para $k \in \mathbb{N}$. Estes valores correspondem ao zero $z$ da função $f(x) = x^2 - k$ definida em $\mathbb{R}^+$. Equivalentemente, podemos obter $z$ como o ponto fixo de

$$N(x) = x - \frac{f(x)}{f'(x)} = \frac{1}{2}\left(x + \frac{k}{x}\right).$$

Como $N(x) = \frac{1}{2}(1 - kx^{-2})$, temos que $\min_{\mathbb{R}^+} g = \sqrt{k}$. Ora, $|N'(x)| = \frac{1}{2}|1 - kx^{-2}|$ é menor que 1 se $x > \sqrt{k/3}$. Por outro lado, se $0 < x \leq \sqrt{k/3}$ então $N(x) > \sqrt{k/3}$ que corresponde ao caso anterior. I.e. $N$ é contractiva para $x > \sqrt{k/3}$ e como $N(D) \subset D$ com $D = ]\sqrt{k/3}, k[$, existe um único ponto fixo $z = \lim x_n$ de $N$ em $D \ni x_0$.

No caso de $k = 2$, obtemos a tabela seguinte:

| $n$ | $x_n$ |
|---|---|
| 0 | 1 |
| 1 | 1.5 |
| 2 | 1.41666 |
| 3 | 1.42422 |
| 4 | 1.41421 |

# 5 Integração numérica

Seja $f\colon [a, b] \to \mathbb{R}$ integrável. O objectivo deste capítulo é o de calcular numericamente

$$I(f) = \int_a^b f.$$

Tanto a função $f$ como a sua primitiva podem ser desconhecidas, daí aproximarmos $I(f)$ tendo em conta apenas os valores de $f$ nalguns pontos.

Dados pontos $x_0, \ldots, x_n \in [a, b]$ onde a função é conhecida (nós de quadratura) e coeficientes $A_0, \ldots, A_n \in \mathbb{R}$ (pesos de quadratura), a **quadratura de** $f$ é dada por:

$$S(f) = \sum_{i=0}^{n} A_i f(x_i).$$

Naturalmente, o erro de aproximação do integral de $f$ pela quadratura é

$$E(f) = S(f) - I(f).$$

Se $E(f) = 0$ dizemos que $S$ é exacta.

## 5.1 Grau de quadratura

O erro cometido irá estar associado ao **grau de exactidão polinomial** (ou grau de aproximação) de $S$ definido da seguinte forma:

$$\text{grau}(S) = k \quad \text{sse} \quad \begin{cases} S_n(x^j) = I(x^j), & 0 \leq j \leq k, \\ S_n(x^{k+1}) \neq I(x^{k+1}). \end{cases}$$

**Proposição 5.1.** $\text{grau}(S) = k$ *sse*

1. *para qualquer* $P \in \mathcal{P}_k$ *temos que* $S(P) = I(P)$,
2. *existe* $Q \in \mathcal{P}_{k+1}$ *tal que* $S(Q) \neq I(Q)$.

**Exercício 5.2.** *Prove a proposição anterior.*

**Proposição 5.3.** $\text{grau}(S_n) \leq 2n + 1$.

*Demonstração.* Considere o polinómio $N(x) = \prod_{i=0}^{n}(x - x_i)$ com grau $n+1$. Então $N^2$ é um polinómio de grau $2n + 2$ com $n + 1$ zeros $x_i$. Assim, $S(N^2) = \sum_i A_i N(x_i)^2 = 0$. Como $I(N^2) > 0$ temos que ter $S(N^2) \neq I(N^2)$. Ou seja, $\text{grau}(S) < 2n + 2$. □

## 5.2 Exemplos de quadraturas

Observe que usando a aproximação da função $f$ dada pelo seu polinómio interpolador $P = \sum_i f(x_i)\, L_i$, o integral de $f$ deverá ser aproximado por

$$I(P) = \sum_{i=0}^{n} f(x_i) I(L_i).$$

Usando os pesos dados por $A_i = I(L_i)$ esta expressão corresponde a uma quadratura de $f$.

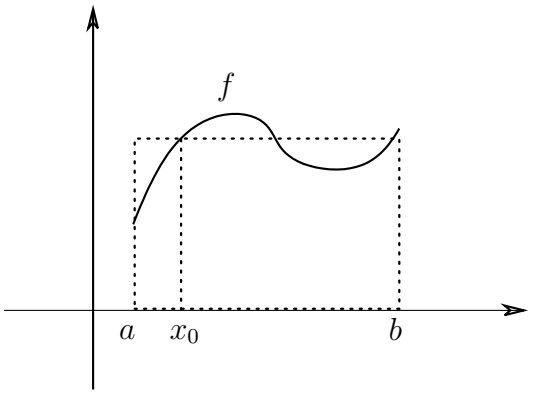


Figura 4: Exemplo da quadratura do rectângulo.

**Proposição 5.4.** *Se* $A_i = I(L_i)$*, então* $\text{grau}(S) \geq n$.

*Demonstração.* Se $j \leq n$, o polinómio interpolador de $x^j$ é $P(x) = x^j$. Logo, $S(x^j) = I(x^j)$. □

Vamos considerar apenas quadraturas com pesos dados por $A_i = I(L_i)$. Assim,

$$n \leq \text{grau}(S) \leq 2n + 1.$$

Seguem-se alguns exemplos, versões para diferente número ou escolhas específicas de nós de quadratura.

### 5.2.1 Quadratura do rectângulo ($n = 0$)

Considere apenas um nó $x_0 \in [a, b]$ e tome $A_0 = I(L_0) = \int_a^b 1 = b - a$. Assim,

$$S(f) = A_0 f(x_0) = (b - a) f(x_0).$$

Ou seja, escolhendo $x_0$ a quadratura dá-nos a área do rectângulo com base $[a, b]$ e altura $f(x_0)$ (ver Figura 4).

Pela Proposição 5.4 temos $0 \leq \text{grau}(S) \leq 1$. Para verificar o grau de $S$ basta calcular $S(x) = A_0 x_0 = (b - a)x_0$ e $I(x) = (b^2 - a^2)/2$. Portanto, $\text{grau}(S) = 1$ sse $S(x) = I(x)$ sse $x_0 = (b + a)/2$.

Concluimos assim que $\text{grau}(S) = 0$ excepto quando escolhemos $x_0$ como o ponto médio do intervalo $[a, b]$. Este caso particular da quadratura do rectângulo chama-se **quadratura do ponto médio**.

**Exemplo 5.5.** Queremos obter aproximações pela quadratura do rectângulo do valor de $\int_0^\pi \sin(x)\, dx$. Pela descrição anterior, $S(\sin) = \pi \sin(x_0)$. Se escolhermos $x_0 = \pi/2$ (quadratura do ponto médio) obtemos $S(\sin) = 3.1415...$. Note que o valor exacto é $I(\sin) = 2$.

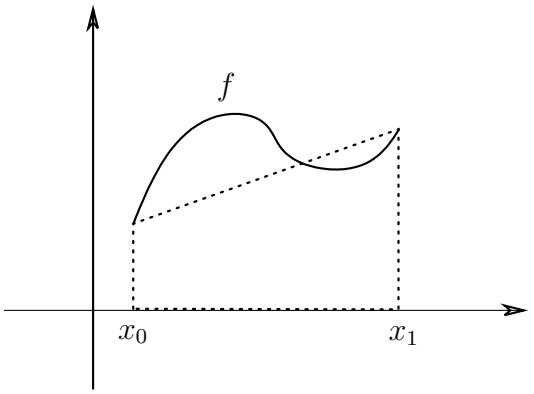


Figura 5: Exemplo da quadratura do trapézio.

### 5.2.2 Quadratura do trapézio ($n = 1$)

Neste caso consideramos $x_0 = a$, $x_1 = b$ e

$$A_0 = I(L_0) = \int_{x_0}^{x_1} \frac{x - x_1}{x_0 - x_1} dx = \frac{x_1 - x_0}{2}$$
$$A_1 = I(L_1) = A_0.$$

Logo,

$$S(f) = A_0 f(x_0) + A_1 f(x_1) = (x_1 - x_0)\frac{f(x_1) + f(x_0)}{2}.$$

Este valor corresponde à área do trapézio definido pela base $[x_0, x_1]$ e limitada pelo segmento que une $(x_0, f(x_0))$ a $(x_1, f(x_1))$ (ver Figura 5).

Agora $1 \leq \text{grau}(S) \leq 3$. Escrevendo $S(x^2) = (x_0^2 + x_1^2)(x_1 + x_0)/2$ é simples verificar que é diferente de $I(x^2) = (x_1^3 - x_0^3)/3$. Então $\text{grau}(S) = 1$.

**Exemplo 5.6.** A aproximação pela quadratura do trapézio de $\int_0^\pi \sin(x)\, dx$ com os valores nodais $\sin(0) = 0$ e $\sin(\pi) = 0$ e com $A_0 = A_1 = \pi/2$, é dada por $S(\sin) = 0$.

### 5.2.3 Quadratura de Simpson ($n = 2$)

Aumentando o número de nós para três: $x_0 = a$, $x_1$ e $x_2 = b$, e escolhendo $A_i = I(L_i)$ novamente, obtemos a quadratura de Simpson (ver Figura 6):

$$S(f) = A_0 f(x_0) + A_1 f(x_1) + A_2 f(x_2).$$

Os pontos $x_i$ são interpolados por um polinómio de grau $\leq 2$ que define uma região cuja área aproxima o integral de $f$.

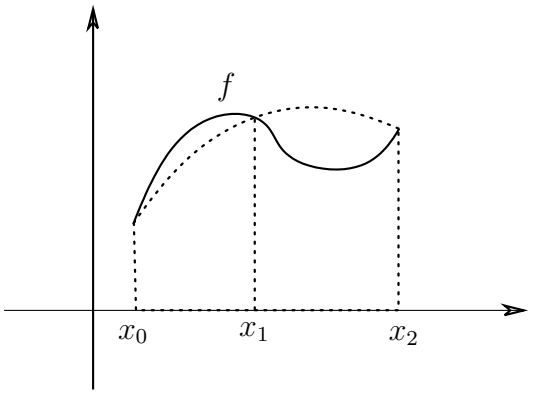


Figura 6: Exemplo da quadratura de Simpson.

**Exemplo 5.7.** Usando a quadratura de Simpson para calcular numericamente $\int_0^\pi \sin(x)\,dx$ nos nós $x_0 = 0, x_1, x_2 = \pi$, e sabendo que $\sin(0) = \sin(\pi) = 0$, basta calcular

$$A_1 = I(L_1) = \int_0^\pi \frac{x(x-\pi)}{x_1(x_1-\pi)}dx = \frac{\pi^3}{6x_1(\pi - x_1)}.$$

Finalmente, $S(\sin) = A_1 \sin(x_1)$. Se escolhermos $x_1 = \pi/2$ obtemos $S(\sin) = 2.094....$

**Exercício 5.8.** *Determine a fórmula geral explícita da quadratura de Simpson (i.e. para $n = 2$ determine $A_i = I(L_i)$ em função de $a = x_0$, $x_1$ e $b = x_2$), e calcule o seu grau.*

### 5.2.4 Quadraturas compostas

Equipados com as quadraturas anteriores, podemos calcular integrais de funções considerando partições do intervalo. Em cada subintervalo assim definido determinamos a quadratura de $f$. O resultado final para a aproximação de $I(f)$ será dado como a soma sobre todos os subintervalos.

A partição é formada por nós de quadratura disponíveis. Cada subintervalo da partição deverá conter os nós necessários para a determinação de todos os pesos de acordo com a quadratura escolhida.

**Exemplo 5.9.** Queremos determinar $\int_0^1 f$ com $f(x) = e^{-x^2}$. Para isso escolhemos a partição do intervalo $[0, 1]$ dada pelos nós $x_j = j/N$, $j = 0, \ldots, N$, com $N \in \mathbb{N}$. Vamos utilizar duas quadraturas: do ponto médio $S_0$ e do trapézio $S_1$.

- Usando a quadratura do ponto médio para cada subintervalo $[x_j, x_{j+1}]$,

temos $A_0 = x_{j+1} - x_j = 1/N$. Logo,

$$S_0(f) = \sum_{j=0}^{N-1} \frac{e^{-\left(\frac{j}{N}+\frac{1}{2N}\right)^2}}{N}.$$

- Para a quadratura do trapézio em cada subintervalo $[x_j, x_{j+1}]$, obtemos

$$S_1(f) = \sum_{j=0}^{N-1} \frac{e^{-\left(\frac{j+1}{N}\right)^2} + e^{-\left(\frac{j}{N}\right)^2}}{2N}.$$

A tabela seguinte dá-nos as aproximações para diversos valores de $N$. Note que o valor exacto é 0.746824....

| $N$ | ponto médio | trapézio |
|---|---|---|
| 1 | 0.778801 | 0.68394 |
| 2 | 0.754598 | 0.73137 |
| 3 | 0.750252 | 0.739986 |
| 4 | 0.748747 | 0.742984 |
| 5 | 0.748053 | 0.744368 |
| 6 | 0.747677 | 0.745119 |
| 7 | 0.747451 | 0.745572 |
| 8 | 0.747304 | 0.745866 |
| 9 | 0.747203 | 0.746067 |
| 10 | 0.747131 | 0.746211 |

**Exercício 5.10.** *Repita o exemplo anterior usando a quadratura de Simpson.*

## 5.3 Erros de quadratura

Como a nossa escolha de pesos foi sempre de $A_i = I(L_i)$, logo $S(f) = I(P)$ para o polinómio interpolador $P$ de $f$. Assim, o erro é dado por

$$E(f) = S(f) - I(f) = I(P - f).$$

Ou seja, o erro é estimado a partir do erro de interpolação. Então, nas condições do Teorema 3.30,

$$|E(f)| \leq (b-a)\frac{\|f^{(n+1)}\|_{C^0}}{(n+1)!}\|N_{n+1}\|_{C^0}.$$

**Exemplo 5.11.** No caso da quadratura do rectângulo onde temos apenas um nó ($n = 0$),

$$|E(f)| \leq (b-a)\|f'\|_{C^0} \max\{x_0 - a, b - x_0\} \leq (b-a)^2\|f'\|_{C^0}.$$

Note que se $f$ é um polinómio de grau 0, então $f' = 0$ e a quadratura é exacta (o erro é nulo). Como seria de esperar pois grau$(S) = 0$.

**Exercício 5.12.** *Estime o erro usando outras quadraturas.*

## 5.4 Quadratura de Gauss

Nesta secção queremos obter a melhor quadratura possível em termos de exactidão polinomial. Isto é, queremos determinar condições nos nós para os quais tenhamos quadraturas de grau máximo possível: $2n+1$. Em diversas aplicações este método não pode ser utilizado por não termos liberdade de escolha dos nós. Iremos usar os pesos $A_i = I(L_i\rho)$ para uma função densidade $\rho \geq 0$ fixa à partida tal que

$$\int_a^b \rho(x)\,dx > 0.$$

Isto permitirá aproximar integrais mais gerais do tipo

$$I(f) = \int_a^b f(x)\,\rho(x)\,dx.$$

Como anteriormente, os nós $x_i$, $i = 0, \ldots, n$, da quadratura definem inequivocamente o polinómio $N_{n+1}$ de grau $n+1$.

### 5.4.1 Produto interno em $\mathcal{P}_n$

Vamos recordar a noção de produto interno num espaço vectorial $\mathcal{V}$ sobre os reais. Uma função $\langle\cdot,\cdot\rangle\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R}$ é um produto interno se verificar as seguintes condições para quaisquer $u, u', v \in \mathcal{V}$, $\alpha \in \mathbb{R}$:

1. $\langle u, u\rangle > 0$
2. $\langle \alpha u, v\rangle = \alpha\langle u, v\rangle$
3. $\langle u + u', v\rangle = \langle u, v\rangle + \langle u', v\rangle$
4. $\langle u, v\rangle = \langle v, u\rangle$

Considere o espaço $L^2([a,b])$ das funções $f$ de quadrado integrável (relativamente à medida de Lebesgue) em $[a,b]$, i.e. $\int_a^b f(x)^2\,dx < +\infty$. Verifica-se facilmente que para qualquer escolha de uma função $\rho\colon [a,b]\to\mathbb{R}_0^+$ em $L^2$,

$$\langle f, g\rangle = I(fg\rho) = \int_a^b f(x)\,g(x)\,\rho(x)\,dx$$

é um produto interno. Qualquer polinómio restringido a $[a,b]$ pertence obviamente a $L^2([a,b])$.

**Exemplo 5.13.** A função $f(x) = x^{-1/2}$ é integrável em $[0,1]$, mas não pertence a $L^2([0,1])$ pois $x^{-1}$não é integrável nesse intervalo.

**Exercício 5.14.** *Mostre que qualquer função de quadrado integrável é integrável.*

Se $\langle f, g\rangle = 0$ dizemos que $f$ e $g$ são ortogonais e escrevemos $f \perp g$. Uma função $f$ é ortogonal a $\mathcal{P}_n$ (i.e. $f \perp \mathcal{P}_n$) sse $f$ é ortogonal a todos os polinómios em $\mathcal{P}_n$.

**Teorema 5.15.** *$N_{n+1} \perp \mathcal{P}_n$ sse* $\text{grau}(S) = 2n + 1$.

**Observação 5.16.** O teorema anterior apesar de compacto contém muita informação. Significa que podemos obter uma quadratura de grau máximo sse existirem nós, zeros de $N_{n+1}$, tais que $\langle N_{n+1}, q\rangle = 0$ para qualquer $q \in \mathcal{P}_n$. Sucessões deste tipo de polinómios $N_{n+1}$ existem (como iremos ver mais adiante) e têm $n + 1$ raízes $x_i$ usadas como os nós da quadratura.

*Demonstração.* ($\Leftarrow$) Assumindo que $\text{grau}(S) = 2n + 1$ e $q \in \mathcal{P}_n$, temos que $I(N_{n+1}q) = S_n(W_n q)$ pois $N_{n+1}q$ é um polinómio de grau $\leq 2n + 1$. Finalmente, $S(N_{n+1}q) = \sum_i A_i N_{n+1}(x_i)q(x_i) = 0$.

($\Rightarrow$) Temos que $n \leq \text{grau}(S) \leq 2n + 1$. É suficiente provar que $I(x^j) = S(x^j)$ para qualquer $n + 1 \leq j \leq 2n + 1$. Ora, pelo algoritmo de divisão de polinómios,

$$x^j = Q(x)N_{n+1}(x) + R(x),$$

onde $Q$ tem grau $j - (n + 1) \leq n$ e $R$ tem grau $\leq n$, i.e. $Q, R \in \mathcal{P}_n$. Logo, $x_i^j = R(x_i)$ pois $N_{n+1}(x_i) = 0$. Por outro lado, $I(x^j) = I(QN_{n+1}) + I(R) = I(R)$ porque $N_{n+1} \perp Q$. Também temos que $I(R) = S(R)$ pois $\text{grau}(S) \geq n$. Juntando todos os resultados anteriores,

$$I(x^j) = \sum_i A_i R(x_i) = \sum_i A_i x_i^j = S(x^j),$$

como queriamos demonstrar. □

### 5.4.2 Escolha dos nós

Vamos agora encontrar os nós que definem o polinómio $N_{n+1}$ tornando-o ortogonal a $\mathcal{P}_n$. Estes nós irão corresponder aos zeros de polinómios com determinadas propriedades como iremos ver de seguida.

Um conjunto de polinómios $\{p_0, \ldots, p_n\}$ é um **sistema ortogonal** de $\mathcal{P}_n$ em $[a, b]$ sse

1. $\text{grau}(p_i) = i$,
2. $\langle p_i, p_j\rangle = 0$, $i \neq j$.

**Observação 5.17.** Como o grau de cada $p_i$ é igual a $i$, um sistema ortogonal forma uma base ortogonal de $\mathcal{P}_n$.

**Exemplo 5.18.** O processo de ortogonalização de Gram-Schmidt permite-nos construir um sistema ortogonal. Considere a base $\{1, x, \ldots, x^n\}$ de $\mathcal{P}_n$ em $[-1, 1]$ e $\rho = 1$. Esta base não é ortogonal pois e.g. $\langle 1, x^2\rangle = \int_{-1}^{1} x^2\, dx =$

$2/3 \neq 0$. Podemos contudo construir uma base ortogonal a partir desta. Escolhemos $p_0(x) = 1$ e os restantes elementos da base são dados por

$$p_i(x) = x^i - \sum_{j=0}^{i-1} \langle x^i, p_j \rangle \, p_j(x), \qquad i = 1, \ldots, n.$$

A partir do facto

$$\int_{-1}^{1} x^k dx = \begin{cases} 0, & k \text{ ímpar} \\ \frac{2}{j+1}, & k \text{ par,} \end{cases}$$

obtemos o sistema ortogonal com os primeiros elementos: $p_1(x) = x$ e $p_2(x) = x^2 - \frac{2}{3}$, etc.

**Proposição 5.19.** *Seja $\{p_0, \ldots, p_n\}$ um sistema ortogonal de $\mathcal{P}_n$ em $[a, b]$. Então, para cada $0 \leq i \leq n$,*

1. *$p_i \perp \mathcal{P}_j$, $0 \leq j \leq i - 1$,*
2. *$p_i$ tem $i$ zeros reais e distintos em $[a, b]$,*
3. *existem $a, b, c \in \mathbb{R}$ tais que*

$$p_{i+1}(x) = (ax + b)p_{i+1}(x) + cp_i(x).$$

*Demonstração.*

1. Qualquer $P \in \mathcal{P}_j$ pode ser escrito como uma combinação linear dos vectores da base ortogonal $\{p_0, \ldots, p_j\}$, $P = \sum_{k=0}^{j} c_k p_k$. Logo $\langle P, p_i \rangle = \sum_k c_k \langle p_k, p_i \rangle = 0$.

2. Seja $0 \leq m \leq i$ o número de zeros de $p_i$, denotados por $z_1, \ldots, z_m$, que verificam as condições seguintes:

   - $z_j \in [a, b]$,
   - $p_i'(z_j) \neq 0$ (i.e. o sinal de $p_i$ muda em cada $z_j$).

   Então $p_i(x) = Q(x) \prod_{j=1}^{m} (x - z_j)^{r_j}$ onde $Q$ é um polinómio de grau $i - \sum_{j=1}^{m} r_j$ que não muda de sinal (é $\geq 0$ ou $\leq 0$) e cada $r_j$ é ímpar (se $r_j$ fosse par $p_i$ não mudava de sinal em $z_j$).

   Suponha que $m < i$. Tomando o polinómio em $\mathcal{P}_m$ dado por $V(x) = \prod_{i=1}^{m} (x - z_i)$ temos que

$$\langle V, p_i \rangle = \int_a^b Q(x) \prod_{j=1}^{m} (x - z_j)^{r_j + 1} \rho(x) \, dx.$$

   Agora $r_j + 1$ é par, logo a função integranda tem o mesmo sinal de $Q$ que não muda. Além disso, como os polinómios só assumem o valor

zero num número finito de pontos, o integral só seria nulo se $\rho = 0$ q.t.p. O que não se verifica pois $\int \rho > 0$. Então, $\langle V, p_i \rangle \neq 0$. Ou seja, existe um polinómio em $\mathcal{P}_m$ que não é ortogonal a $p_i$, logo $p_i$ não pode pertencer a um sistema ortogonal. Mostrámos assim que para um sistema ortogonal temos que ter $m = i$. Ou seja, $p_i$ tem grau $i$ e $i$ zeros em $[a, b]$ que são distintos.

3. Escolhemos $a$ tal que $p_{i+2}(x) - axp_{i+1}(x)$ tem grau $i + 1$. Escolhemos $b$ tal que $q(x) = p_{i+2}(x) - (ax + b)p_{i+1}(x)$ tem grau $i$. Assim, $q(x) = \sum_{j=0}^{i} \alpha_j p_j$. Além disso, para $k = 0, \ldots, i$ temos $0 = \langle q, p_k \rangle = \alpha_k \langle p_k, p_k \rangle$. Ou seja, $\alpha_k = 0$. Resta assim o termo $q = \alpha_i p_i$, i.e. $p_{i+2}(x) - (ax + b)p_{i+1}(x) = \alpha_i p_i$.

□

A proposição anterior permite-nos concluir que a escolha dos nós de uma quadratura com grau máximo pode ser feita a partir dos zeros dos polinómios de um sistema ortogonal. Iremos de seguida apresentar exemplos de sistemas ortogonais e calcular os zeros de alguns dos seus polinómios.

### 5.4.3 Polinómios de Legendre em $[-1, 1]$

Considerando $P_0(x) = 1$ e $P_1(x) = x$ definimos

$$P_n(x) = \frac{2n-1}{n} x P_{n-1}(x) - \frac{n-1}{n} P_{n-2}(x), \qquad n \geq 2.$$

**Exercício 5.20.**

1. *Prove que* $\text{grau}(P_n) = n$, $P_n(\pm 1) = (\pm 1)^n$, $P_n$ *é uma função par para* $n$ *par e ímpar para* $n$ *ímpar.*

2. *Defina os conjuntos* $Z_n$ *dos zeros de* $P_n$. *Mostre que*

$$Z_0 = \emptyset$$
$$Z_1 = \{0\}$$
$$Z_2 = \left\{ -\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3} \right\}$$
$$Z_3 = \left\{ -\sqrt{\frac{3}{5}}, 0, \sqrt{\frac{3}{5}} \right\}$$
$$Z_4 = \left\{ -\sqrt{\frac{3}{7} + \frac{2\sqrt{30}}{35}}, -\sqrt{\frac{3}{7} - \frac{2\sqrt{30}}{35}}, \sqrt{\frac{3}{7} - \frac{2\sqrt{30}}{35}}, \sqrt{\frac{3}{7} + \frac{2\sqrt{30}}{35}} \right\}$$
$$Z_5 = \left\{ -\sqrt{\frac{35 + 2\sqrt{70}}{63}}, -\sqrt{\frac{35 - 2\sqrt{70}}{63}}, 0, \sqrt{\frac{35 - 2\sqrt{70}}{63}}, \sqrt{\frac{35 + 2\sqrt{70}}{63}} \right\}.$$

**Exemplo 5.21.** Queremos aproximar $\int_0^\pi \sin(x)\,dx$. Para isso vamos calcular os nós para a quadratura de Gauss como os zeros dos polinómios de Legendre após a transformação de variável para o intervalo $[0,\pi]$ dada por

$$h\colon [-1,1] \to [0,\pi], \qquad h(x) = \frac{\pi}{2}(x+1).$$

Os novos polinómios de Legendre são

$$\widetilde{P}_n = P_n \circ h^{-1}$$

definidos em $[0,\pi]$. Em particular, os zeros de $\widetilde{P}_n$ são os elementos do conjunto $h(Z_n)$.

1. Caso $n = 0$, usando um nó apenas (o zero de $\widetilde{P}_1$): $x_0 = \pi/2$. Pela fórmula da quadratura do rectângulo neste nó, temos

   $$S_0(\sin) = A_0 \sin(x_0) = \pi = 3.1415\ldots.$$

   Esta quadratura de Gauss tem grau 1 e corresponde ao caso da quadratura do ponto médio.

2. Caso $n = 1$, dois nós. Usando a fórmula da quadratura do trapézio,

   $$S_1(\sin) = (x_1 - x_0)\frac{\sin(x_0) + \sin(x_1)}{2} = 1.11765\ldots.$$

3. Caso $n = 2$, usando três nós na fórmula da quadratura de Simpson,

   $$S_2(\sin) = 1.90355\ldots$$

**Exercício 5.22.** *Efectue todos os cálculos do exemplo anterior, completando-o para os casos $n = 3, 4$.*

### 5.4.4 Polinómios de Chebyshev em $[-1,1]$

Considerando $T_0(x) = 1$ e $T_1(x) = x$ definimos

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \qquad n \geq 2.$$

Estes polinómios formam um sistema ortogonal para o produto interno definido com densidade

$$\rho(x) = \frac{1}{\sqrt{1-x^2}}.$$

Usando os pesos

$$A_i = \langle L_i, 1\rangle = \int_{-1}^{1} L_i(x)\frac{dx}{\sqrt{1-x^2}},$$

a quadratura é apropriada para aproximar integrais do tipo

$$\int_{-1}^{1} f(x)\frac{dx}{\sqrt{1-x^2}}.$$

Observe que $\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} = \pi$.

**Exercício 5.23.** *Calcule $A_i$.*

### 5.4.5 Polinómios de Hermite em $\mathbb{R}$

Considerando $H_0(x) = 1$ e $H_1(x) = 2x$ definimos

$$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x), \qquad n \geq 2.$$

Estes polinómios formam um sistema ortogonal para o produto interno com densidade

$$\rho(x) = e^{-x^2}.$$

A sua utilidade reside na aproximação de integrais do tipo

$$\int_{\mathbb{R}} f(x)e^{-x^2}dx.$$

Observe que $\int_{\mathbb{R}} e^{-x^2}dx = \sqrt{\pi}$.

**Exercício 5.24.** *Calcule os zeros de $H_2$ e $H_3$. Use-os para calcular aproximações de $\int_{\mathbb{R}} \sin(x^2)e^{-x^2}\,dx$.*

# 6 Métodos numéricos para edo's

Uma função $f\colon \mathbb{R}^d \times [a,b] \to \mathbb{R}^d$ e $x_0 \in \mathbb{R}^d$ definem a seguinte equação diferencial ordinária multidimensional de 1ª ordem e condição inicial:

$$\dot{x} = f(x,t), \quad t \in [a,b], \quad \text{com} \quad x(a) = x_0,$$

respectivamente. A solução deste problema é a função $x\colon [a,b] \to \mathbb{R}^d$ na forma

$$x(t) = x_0 + \int_a^t f(x(s),s)\,ds. \tag{6.1}$$

Por vezes a solução é também chamada de órbita, trajectória, caminho, fluxo, etc. Como a solução não é simples de obter para a generalidade dos casos em que $f$ não é uma função linear, o nosso objectivo é determinar uma aproximação numérica de $x(t)$.

**Observação 6.1.**

- Pelo teorema de existência e unicidade de solução para edo's, se $x \mapsto f(x,t)$ é Lipschitz[14] e $t \mapsto f(x,t)$ é $C^0$, então existe uma única solução $x\colon [a,b] \to \mathbb{R}^d$.
- Se $d \geq 2$ estamos a trabalhar com vectores. Ou seja, $x(t) = (x_1(t), \ldots, x_d(t))$ e $f(x,t) = (f_1(x(t),t), \ldots, f_d(x(t),t))$. Desta forma, $\int_a^t f(x(s),s)\,ds = (\int_a^t f_1(x(s),s)\,ds, \ldots, \int_a^t f_d(x(s),s)\,ds)$.

[14] $g$ é Lipschitz sse existe $L > 0$ tal que para quaisquer $x_1, x_2$ temos que $\|g(x_1) - g(x_2)\| \leq L\|x_1 - x_2\|$.

**Exercício 6.2.** *Mostre que uma função $C^1$ é Lipschitz, que por sua vez é $C^0$.*

Para aproximarmos o integral em (6.1) temos que considerar nós no intervalo $[a, b]$, correspondendo a uma partição de $[a, b]$ dada por $\{t_0, \ldots, t_N\}$ com $N \in \mathbb{N}$, satisfazendo

$$a = t_0 < t_1 < \cdots < t_N = b.$$

Cada $t_i$ é o $i$-nésimo nó e a diferença $\Delta t_i = t_i - t_{i-1}$ é o passo $i$. Iremos utilizar métodos de passo constante, i.e. $\Delta t_i = h > 0$ para qualquer $i = 1, \ldots, N$.

O nosso objectivo é então determinar valores nodais $x_i$ que aproximem em $t = t_i$ a solução exacta $x(t_i)$. O erro no passo $i$ será naturalmente dado por

$$e_i = x(t_i) - x_i.$$

Usando (6.1) é simples verificar que, com $i = 1, \ldots, N$,

$$\begin{aligned} x(t_i) &= x_0 + \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} f(x(s), s)\, ds \\ &= x(t_{i-1}) + \int_{t_{i-1}}^{t_i} f(x(s), s)\, ds. \end{aligned}$$

Aproximando cada $x(t_i)$ por $x_i$ e o integral acima por $S^{(i)}$ temos então a seguinte fórmula geral de recorrência para as aproximações nos nós:

$$\begin{cases} x_i = x_{i-1} + S^{(i)}, & i = 1, \ldots, N \\ x_0 = x(t_0). \end{cases}$$

As escolhas para $S^{(i)}$ são vastas, cada correspondendo a um método diferente. Apresentamos vários exemplos de métodos nas secções seguintes.

**Observação 6.3.** Note que $S^{(i)}$ não pode ser obtido recorrendo directamente a quadraturas. Isto porque a função integranda $f(x(s), s)$ depende dos valores de $x$ no intervalo $[t_{i-1}, t_i]$ que são desconhecidos.

## 6.1 Erro e ordem do método

Para cada método, i.e. escolha de $S^{(i)}$, o erro pode ser estimado em termos do parâmetro $h = \max_i \Delta t_i$. Vamos escolher sempre uma partição uniforme com $h = \Delta t_i$ para qualquer $i = 1, \ldots, N$. Logo,

$$N = \frac{b - a}{h}.$$

De forma a simplificar as estimativas do erro, supomos que $e_0 = 0$ desprezando o erro de representação numérica.

A ordem de convergência de um método é definida por

$$p = \lim_{h \to 0^+} \frac{\log \|e_N\|}{\log h}.$$

Se $p > 0$, então o método converge para a solução exacta quando $h \to 0^+$. A velocidade de convergência do método é assim determinada por $p$. Note porém que não é possível computacionalmente fazer escolhas de $h$ para além do limite mínimo de representatividade numérica.

Dizemos que uma função $\psi$ é de ordem $h^p$, i.e. $\psi = \mathcal{O}(h^p)$, sse existe $\lim_{h\to 0} h^{-p}\|\psi(h)\|$ finito. Isto significa que $\|\psi\|$ converge para zero como $h^q$ para algum $q \geq p$.

**Exercício 6.4.** *Mostre que:*

1. $\mathcal{O}(h^p) + \mathcal{O}(h^q) = \mathcal{O}(h^{\min\{p,q\}})$.
2. $\mathcal{O}(h^p)\mathcal{O}(h^q) = \mathcal{O}(h^{pq})$.
3. *Se* $\psi \leq \mathcal{O}(h^p)$, *então* $\psi = \mathcal{O}(h^p)$.
4. *Se* $\psi = \mathcal{O}(h^p)$, *então* $\psi = \mathcal{O}(h^q)$ *para qualquer* $q \leq p$.

**Exercício 6.5.** *Mostre que:*

1. *Se um método tem ordem* $p$, *então* $e_N = \mathcal{O}(h^p)$.
2. *Se* $e_N = \mathcal{O}(h^q)$, *então o método tem ordem* $p \geq q$.

Segue-se um critério para determinação da ordem.

**Proposição 6.6.** *Se para quaisquer* $i \in \{1, \ldots, N\}$ *e* $t \in [t_{i-1}, t_{i-1} + h]$

$$f(x(t), t) - h^{-1}S^{(i)} = \mathcal{O}(h^q) + \mathcal{O}(\|e_{i-1}\|),$$

*então o método tem ordem* $p \geq q$.

*Demonstração.* Note que

$$\begin{aligned} e_i - e_{i-1} =& x(t_i) - x(t_{i-1}) - (x_i - x_{i-1}) \\ =& \int_{t_{i-1}}^{t_{i-1}+h} f(x(s), s)\, ds - S^{(i)} \\ =& \int_{t_{i-1}}^{t_{i-1}+h} \left( f(x(s), s) - h^{-1}S^{(i)} \right) ds. \end{aligned}$$

Logo,

$$\|e_i - e_{i-1}\| \leq h \max_{t \in [t_{i-1}, t_{i-1}+h]} \left\| f(x(t), t) - h^{-1}S^{(i)} \right\|.$$

Usando esta relação e também $j \leq N = (b-a)/h$,

$$\begin{aligned}\|e_j\| &\leq \sum_{i=1}^{j} \|e_i - e_{i-1}\| \\ &\leq jh \max_{i=1,\ldots,j} \max_{t\in[t_{i-1},t_{i-1}+h]} \left\| f(x(t),t) - h^{-1}S^{(i)} \right\| \\ &\leq \mathcal{O}(h^q) + \max_{i=1,\ldots,j} \mathcal{O}(\|e_{i-1}\|).\end{aligned}$$

Iniciando com $e_0 = 0$ obtemos que $e_N = \mathcal{O}(h^q)$. □

## 6.2 Exemplos de métodos

Apresentamos de seguida duas das escolhas mais comuns para $S^{(i)}$ pela sua facilidade de implementação numérica.

### 6.2.1 Método de Euler

Baseado na quadratura do rectângulo escolhemos

$$S^{(i)} = (t_i - t_{i-1}) f(x_{i-1}, t_{i-1}).$$

Para um passo constante $\Delta t_i = h$, a fórmula de recorrência é dada por

$$x_i = x_{i-1} + hf(x_{i-1}, t_{i-1}), \qquad x_0 = x(t_0).$$

**Proposição 6.7.** *Se $f$ é $C^1$, então o método de Euler tem ordem $p \geq 1$.*

*Demonstração.* Seja $F(t) = f(x(t), t)$. Assim,

$$F'(t) = \frac{\partial f}{\partial x}(x(t),t)\dot{x}(t) + \frac{\partial f}{\partial t}(x(t),t)$$

com $\dot{x}(t) = f(x(t),t)$. Então, para $t \in [t_{i-1}, t_{i-1}+h]$,

$$F(t) - F(t_{i-1}) = \mathcal{O}(|t - t_{i-1}|) = \mathcal{O}(h).$$

Por outro lado, $G(x) = f(x, t_{i-1})$ é uma função $C^1$ que satisfaz

$$G(x) - G(x_{i-1}) = \mathcal{O}(\|x - x_{i-1}\|)$$

para $x$ próximo de $x_{i-1}$. Como $G'(x) = \frac{\partial f}{\partial x}(x,t)$ e $e_{i-1} = x(t_{i-1}) - x_{i-1}$ temos que

$$F(t_{i-1}) - G(x_{i-1}) = f(x(t_{i-1}), t_{i-1}) - f(x_{i-1}, t_{i-1}) = \mathcal{O}(\|e_{i-1}\|).$$

Finalmente, para $t \in [t_{i-1}, t_{i-1}+h]$,

$$\begin{aligned}\|f(x(t),t) - h^{-1}S^{(i)}\| &= \|f(x(t),t) - f(x_{i-1},t_{i-1})\| \\ &= \|F(t) - G(x_{i-1})\| \\ &\leq \|F(t) - F(t_{i-1})\| + \|F(t_{i-1}) - G(x_{i-1})\| \\ &= \mathcal{O}(h) + \mathcal{O}(\|e_{i-1}\|).\end{aligned}$$

Basta agora usar a Proposição 6.6. □

**Exemplo 6.8.**

1. Considere o problema de valor inicial: $\dot{x} = x$ e $x(0) = 1$. Utilizando um passo constante $\Delta t_i = h$ para o intervalo $[0, 1]$, o método de Euler dá-nos a aproximação:
$$x_i = x_0(1 + h)^i$$
à solução exacta $x(t_i) = e^{t_i} = e^{hi}$. O erro em $t_N = 1$ é então $|e_N| = |e^1 - (1+h)^{1/h}|$ onde usámos $N = 1/h$. É simples verificar que quando $1/h \to +\infty$ (ou seja, $h \to 0$) o erro vai também para zero.

2. Para o sistema de equações diferenciais lineares
$$\begin{cases} \dot{x} = -y \\ \dot{y} = x \end{cases} \quad \text{com} \quad \begin{cases} x(0) = 1 \\ y(0) = 0 \end{cases}$$
o método de Euler dá-nos a fórmula:
$$\begin{cases} x_i = x_{i-1} - hy_{i-1} \\ y_i = y_{i-1} + hx_{i-1}. \end{cases}$$
A solução exacta é $(x(t), y(t)) = (\cos t, \sin t)$.

### 6.2.2 Métodos de Runge-Kutta

Os métodos de Runge-Kutta distinguem-se do anterior pelos termos de ordens superiores. A fórmula geral da escolha dos $S^{(i)}$ é a seguinte. Dados parâmetros $q \in \mathbb{N}$, $a_{j,k}$ com $1 \leq j \leq q$ e $1 \leq k \leq j-1$, $b_j$ com $1 \leq j \leq q$, e $c_j$ com $2 \leq j \leq q$, seja
$$S^{(i)} = \sum_{j=1}^{q} hb_j F_j,$$
onde
$$F_1 = f(x_{i-1}, t_{i-1}), \qquad F_j = f\left(x_{i-1} + \sum_{k=1}^{j-1} ha_{j,k}F_k, t_{i-1} + hc_j\right), \quad j = 2, \ldots, q.$$

Note que para cada escolha dos parâmetros acima obtemos um método diferente. Nas secções seguintes iremos apresentar em mais detalhe os métodos com $q = 1, 2, 3$ e 4, determinando a ordem do método respectivo. Iremos ver que essa implica certas restrições aos parâmetros $a_{j,k}$, $b_j$ e $c_j$. Podemos desde já obter uma dessas relações se pretendermos que a ordem seja pelo menos 1.

**Proposição 6.9.** *Se $\sum_{j=1}^{q} b_j = 1$ e $f \in C^1$, então a ordem do método de Runge-Kutta é $p \geq 1$.*

*Demonstração.* Note que $F_j = F_1 + \mathcal{O}(h)$. Como já visto anteriormente,

$$f(x(t),t) \quad = \quad F_1 + \mathcal{O}(\|x(t_{i-1}) - x_{i-1}\|) + \mathcal{O}(t - t_{i-1}).$$

Assim,

$$\begin{aligned} \|f(x(t),t) - h^{-1}S^{(i)}\| &= \left\| f(x(t),t) - \sum_{j=1}^{q} b_j F_j \right\| \\ &\leq \left\| F_1 - F_1 \sum_{j=1}^{q} b_j \right\| + \mathcal{O}(h) + \mathcal{O}(\|e_{i-1}\|) \\ &= \mathcal{O}(h) + \mathcal{O}(\|e_{i-1}\|). \end{aligned}$$

Resta aplicar a Proposição 6.6. □

### 6.2.3 Método de Runge-Kutta 1ªordem

Escolhendo $q = 1$ temos $b_1 = 1$. Assim, o método que resulta corresponde ao método de Euler.

### 6.2.4 Método de Runge-Kutta 2ªordem

Para $q = 2$ existem várias escolhas dependendo dos parâmetros. O método é dado por

$$x_i = x_{i-1} + hb_1 f(x_{i-1}, t_{i-1}) + hb_2 f(x_{i-1} + ha_{21} f(x_{i-1}, t_{i-1}), t_{i-1} + hc_2).$$

Note que $b_1 = 1 - b_2$. Os restantes parâmetros $b_2$, $a_{21}$ e $c_2$ são determinados de forma que o método seja de ordem 2.

**Proposição 6.10.** *Se $q = 2$, $b_1 + b_2 = 1$ e $b_2 a_{21} = b_2 c_2 = 1/2$, então a ordem é $p = 2$.*

**Exercício 6.11.** *Demonstre a proposição anterior.*

Exemplos comuns de escolha de parâmetros são:

$$\begin{cases} b_1 = b_2 = \frac{1}{2} \\ a_{21} = c_2 = 1, \end{cases} \qquad \begin{cases} b_1 = 0 \\ b_2 = 1 \\ a_{21} = c_2 = \frac{1}{2}. \end{cases}$$

**Exercício 6.12.** *Use o método de RK2 para determinar aproximações das soluções do sistema*

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = x(\rho - z) - y \\ \dot{z} = xy - \beta z \end{cases}$$

*usando os valores $\sigma = 10$, $\beta = 8/3$ e $\rho = 28$. Note que para qualquer condição inicial $(x_0, y_0, z_0)$ todas as órbitas são atraídas para uma região no espaço (atractor de Lorenz*[15]*).*

### 6.2.5 Método de Runge-Kutta 3ªordem

Considerando $q = 3$ o método mais comum corresponde à escolha de parâmetros de forma a termos:

$$x_i = x_{i-1} + \frac{h}{6}(F_1 + 4F_2 + F_3)$$

com

$$\begin{cases} F_1 = f(x_{i-1}, t_{i-1}) \\ F_2 = f(x_{i-1} + hF_1/2, t_{i-1} + h/2) \\ F_3 = f(x_{i-1} + h(-F_1 + 2F_2), t_{i-1} + h). \end{cases}$$

**Exercício 6.13.** **Para o caso $d = 1$, mostre que a ordem é $p = 3$.*

### 6.2.6 Método de Runge-Kutta 4ªordem

Para $q = 4$ o método mais usado é dado por

$$x_i = x_{i-1} + \frac{h}{6}(F_1 + 2F_2 + 2F_3 + F_4)$$

com

$$\begin{cases} F_1 = f(x_{i-1}, t_{i-1}) \\ F_2 = f(x_{i-1} + hF_1/2, t_{i-1} + h/2) \\ F_3 = f(x_{i-1} + hF_2/2, t_{i-1} + h/2) \\ F_4 = f(x_{i-1} + hF_3, t_{i-1} + h). \end{cases}$$

**Exercício 6.14.** **Para o caso $d = 1$, mostre que a ordem é $p = 4$.*

**Exercício 6.15.** *Escreva um programa para a solução de uma edo de $2^a$ ordem usando o método de Runge-Kutta de $4^a$ ordem. Use-o para determinar a órbita no plano $(x, \dot{x})$ da solução de $\ddot{x} + \sin(x) + 0.01x = 0$ com $x(0) = 0$ e:*

1. $\dot{x}(0) = 1$
2. $\dot{x}(0) = 2.1$
3. $\dot{x}(0) = 3$

**Exercício 6.16.** *Para as seguintes edo's, determine cada ponto de equilíbrio e respectiva estabilidade local. Escreva um programa para a solução numérica usando o método de Runge-Kutta de $4^a$ ordem e apresente o gráfico.*

[15] http://en.wikipedia.org/wiki/Lorenz_attractor

1. *Equações de Rössler:*

$$\begin{cases} \dot{x} = -y - z \\ \dot{y} = x + ay \\ \dot{z} = b + z(x - c) \end{cases}$$

*para os parâmetros* $a = b = 0.1$ *e* $c = 14$. *Observe que também existe uma atractor (atractor de Rössler*[16]*).*

2. *Equações de Lotka-Volterra (sistemas de predador-presa*[17]*):*

$$\begin{cases} \dot{x} = x(\alpha - \beta y) \\ \dot{y} = -y(\gamma - \delta x) \end{cases}$$

*para os parâmetros* $\alpha = \beta = \gamma = \delta = 1$. *Interprete os resultados obtidos tendo em conta que* $x$ *representa a população de presas e* $y$ *a de predadores.*

## Agradecimentos

A Bilal Machraa, Jorge Veloso, Francisco Dias e Carlos Brás do MAEG, pela ajuda na detecção de gralhas em versões anteriores deste texto.

## Referências

[1] H. Pina. *Métodos Numéricos.* McGraw-Hill, 1995.

[2] K.E. Atkinson. *An Introduction to Numerical Analysis.* Wiley, 1989.

[3] R. Burden, J. Faires, A. Burden. *Numerical Analysis.* Cengage Learning, 2014.

[16] http://en.wikipedia.org/wiki/R%C3%B6ssler_map
[17] http://en.wikipedia.org/wiki/Lotka-Volterra_equation