Tópicos de Métodos Numéricos

Gaspar J. Machado

Departamento de Matemática, Universidade do Minho

dezembro de 2024

Índice

1 Erros e estabilidade

2 Equações não lineares

3 Sistemas de equações lineares

4 Interpolação polinomial

5 Quadratura numérica

Obs 1.1

Para efetuarmos cálculos com números reais é necessário começar por escolher um seu sistema de representação.

Def 1.2

⟦representação de um número real na base decimal⟧ Sejam $x \in \mathbb{R}$ e $n \in \mathbb{Z}$. A sequência

$$\sigma \times (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots)$$

com

$$\sigma \in \{-1, +1\},\ d_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}, i = n, n-1, \ldots,$$

diz-se uma representação de x na base decimal se

$$x = \sigma \times \sum_{i=-\infty}^{n} d_i \times 10^i,$$

ou seja, se

$$x = \sigma \times \Big(d_n \times 10^n + d_{n-1} \times 10^{n-1} + \cdots + d_1 \times 10^1 + d_0 \times 10^0 + d_{-1} \times 10^{-1} + d_{-2} \times 10^{-2} + \cdots \Big).$$

Obs 1.3

Na seguinte definição generaliza-se a representação de números reais para qualquer base $\beta \in \mathbb{N} - \{1\}$ (os dois casos mais relevantes são $\beta = 10$, que se diz uma base decimal, e $\beta = 2$, que se diz uma base binária).

Def 1.4

⟦representação de um número real numa base genérica⟧ Sejam $x \in \mathbb{R}$, $n \in \mathbb{Z}$ e $\beta \in \mathbb{N} - \{1\}$. A sequência

$$\sigma \times (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots)$$

com

$$\sigma \in \{-1, +1\},\ d_i \in \{0, \ldots, \beta - 1\}, i = n, n-1, \ldots,$$

diz-se uma representação de x na base β se

$$x = \sigma \times \sum_{i=-\infty}^{n} d_i \times \beta^i.$$

Exe 1.5

Determine o valor do número real x cuja representação na base binária é 101.11.

Res

$$x = 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 1\times 2^{-2} = 4 + 1 + 0.5 + 0.25 = 5.75.$$

Obs 1.6

A necessidade de representar números de diferentes grandezas de uma forma compacta motiva a seguinte definição.

Def 1.7

⟦representação em notação científica de um número real numa base genérica⟧ Sejam $x \in \mathbb{R}$, $n \in \mathbb{Z}$, $t \in \mathbb{Z}$ e $\beta \in \mathbb{N} - \{1\}$. A sequência

$$\sigma \times (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots) \times \beta^t$$

com

$$\sigma \in \{-1, +1\},\ d_i \in \{0, \ldots, \beta - 1\}, i = n, n-1, \ldots,$$

diz-se uma representação de x em notação científica na base decimal se

$$x = \sigma \times \sum_{i=-\infty}^{n} d_i \times \beta^{i+t}.$$

Obs 1.8

Na seguinte definição, por forma a evitar ambiguidades na representação, considera-se a representação normalizada.

Def 1.9

⟦representação em notação científica normalizada de um número real não-nulo, mantissa, expoente⟧ Sejam $x \in \mathbb{R} - \{0\}$, $\beta \in \mathbb{N} - \{1\}$ e $t \in \mathbb{Z}$. A sequência

$$\sigma \times (.d_1 d_2 \cdots) \times \beta^t,$$

com

$$\sigma \in \{-1, +1\},\ d_1 \in \{1, \ldots, \beta-1\},\ d_i \in \{0, 1, \ldots, \beta-1\},\ i = 2, 3, \ldots,$$

diz-se uma representação de x em notação científica normalizada na base β se

$$x = \sigma \times \sum_{i=1}^{+\infty} d_i \times \beta^{-i+t}.$$

A $m = .d_1 d_2 \cdots \in]0, 1[$ chama-se mantissa e a t chama-se expoente.

Exe 1.10

(a) Indique a representação em notação científica normalizada de $x = 123.4567$ na base decimal.

(b) Indique a representação em notação científica normalizada de $x = 0.001234567$ na base decimal.

Res

(a) $x = 0.1234567 \times 10^3$.

(b) $x = 0.1234567 \times 10^{-2}$.

Obs 1.11

(a) Note-se que o número zero não tem uma representação através da notação científica normalizada pois em notação científica normalizada o primeiro dígito da mantissa é sempre diferente de zero.

(b) O facto de, para cálculos em computadores, ser conveniente considerar mantissas com comprimento finito motiva a seguinte definição.

Def 1.12

⟦representação de um número real com precisão finita⟧ Sejam $x \in \mathbb{R}$, $\beta \in \mathbb{N} - \{1\}$ e $t \in \mathbb{Z}$. Uma sequência

$$\sigma \times m \times \beta^t$$

com

$$\sigma \in \{-1, +1\}$$

diz-se uma representação de um número real com precisão finita se o comprimento da mantissa m é finito.

Def 1.13

⟦sistema de vírgula flutuante normalizado⟧ Sejam $\beta \in \mathbb{N} - \{1\}$ e $n \in \mathbb{N}$ e $t_1, t_2 \in \mathbb{N}$ com $t_1 \leqslant t_2$. Chama-se sistema de vírgula flutuante normalizado na base β, comprimento da mantissa n e limites inferior e superior do expoente t_1 e t_2, respetivamente, e que se representa por $\mathbb{F}(\beta, n, t_1, t_2)$ ao conjunto

$$\begin{aligned}\mathbb{F}(\beta, n, t_1, t_2) \overset{\text{def}}{=} \{0\} \cup \big\{ & \sigma \times (.d_1 d_2 \cdots d_n) \times \beta^t : \\ & \sigma \in \{-1, +1\}, \\ & d_1 \in \{1, \ldots, \beta - 1\}, \quad d_i \in \{0, 1, \ldots, \beta - 1\}, i = 2, \ldots, n, \\ & t \in \{t_1, t_1 + 1, \ldots, t_2 - 1, t_2\}\big\}.\end{aligned}$$

Obs 1.14

Note-se que o número zero pertence a qualquer sistema de vírgula flutuante normalizado, embora como um caso particular (veja-se a observação Obs 1.7 (a)).

Exe 1.15

Indique o valor lógico da seguinte proposição: "O número $x = 100$ tem uma representação no sistema de vírgula flutuante normalizado $\mathbb{F}(10, 6, -5, 5)$."

Res

Uma vez que o valor da representação $+0.100000 \times 10^3$, que pertence a $\mathbb{F}(10, 6, -5, 5)$, é 100, a proposição é verdadeira.

Alg 1.16

Conversão de uma representação decimal de um número natural para uma representação binária

input $n \in \mathbb{N}$

output representação binária de n

---

1. Dividir n por 2 e guardar o quociente e o resto.
2. Substituir n pelo quociente obtido na divisão anterior.
3. Repetir os passos 1 e 2 até que o quociente seja 0.
4. A sequência com os restos anotados, lidos de último para o primeiro, é o output.

Exe 1.17

Determine a representação do número na base decimal 18 na base binária.

Res

Atendendo a

$$\begin{aligned} 18 &= 2 \times 9 + 0, \\ 9 &= 2 \times 4 + 1, \\ 4 &= 2 \times 2 + 0, \\ 2 &= 2 \times 1 + 0, \\ 1 &= 2 \times 0 + 1, \end{aligned}$$

tem-se que a representação de 18 na base binária é

$$10010.$$

Confirmação:

$$(10010)_2 = 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 16 + 2 = 18.$$

Alg 1.18

Conversão de uma representação decimal de um número real do intervalo $]0,1[$ para uma representação binária

input $x \in]0,1[$

output representação binária de x

---

1. Multiplicar x por 2.
2. Anotar a parte inteira do resultado (0 ou 1). Esta será a próxima casa decimal do número binário.
3. Substituir o número decimal pelo valor da parte fracionária do resultado.
4. Repetir os passos 1 a 3 até que a parte fracionária se torne 0.

Exe 1.19

Determine a representação de 0.375 na base binária.

Res

Atendendo a

$$2 \times 0.375 = 0.75 = 0 + 0.75,$$
$$2 \times 0.75 = 1.5 = 1 + 0.5,$$
$$2 \times 0.5 = 1 = 1 + 0,$$

tem-se que a representação de 0.125 na base binária é

$$.011$$

Confirmação:

$$(.011)_2 = 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} = 0.25 + 0.125 = 0.375.$$

Exe 1.20

Determine a representação de 0.2 na base binária.

Res

Atendendo a

$$2 \times 0.2 = 0.4 = 0 + 0.4,$$
$$2 \times 0.4 = 0.8 = 0 + 0.8,$$
$$2 \times 0.8 = 1.6 = 1 + 0.6,$$
$$2 \times 0.6 = 1.2 = 1 + 0.2,$$

tem-se que a representação de 0.2 na base binária é

$$.00110011\ldots = .\underline{0011}.$$

Tem-se, então, que o número decimal 0.2 é uma dízima periódica na base binária.

Exe 1.21

Confirme o resultado anterior calculando o valor da série.

Exe 1.22

(a) Indique a representação em notação científica normalizada em base binária do número $x = 1.375$.

(b) Indique o valor lógico da seguinte proposição: "Existe uma representação no sistema de vírgula flutuante normalizado $\mathbb{F}(2,6,-5,5)$ do número $x = 1.375$."

Res

(a) $x = 0.1011 \times 2^1$.

(b) Uma vez que a representação da alínea anterior pertence a $\mathbb{F}(2,6,-5,5)$, a proposição é verdadeira.

Exe 1.23

(a) Indique a representação em notação científica normalizada em base binária do número $x = 0.2$.

(b) Indique o valor lógico da seguinte proposição: "Existe uma representação no sistema de vírgula flutuante normalizado $\mathbb{F}(2, 6, -5, 5)$ do número $x = 0.2$."

Res

(a) $x = 0.\underline{1100} \times 2^{-2}$.

(b) Uma vez que a representação da alínea anterior não pertence a $\mathbb{F}(2, 6, -5, 5)$, a proposição é falsa.

Teo 1.24

(a) Todos os sistemas de vírgula flutuante normalizados são conjuntos finitos.

(b) Todos os sistemas de vírgula flutuante normalizados são conjuntos limitados.

Exe 1.25

Determine a representação em notação científica normalizada em base binária do número $x = 123$.

Exe 1.26

Determine a representação em notação científica normalizada em base binária do número $x = 0.35$.

Exe 1.27

Determine a representação em notação científica normalizada em base binária do número $x = 123.35$.

Exe 1.28

Escreva uma função em MATLAB/Octave cuja variável de entrada seja um número natural em base binária e cuja variável de saída seja a sua representação em base decimal.

Exe 1.29

Escreva uma função em MATLAB/Octave cuja variável de entrada seja um número natural em base decimal e cuja variável de saída seja a sua representação em base binária.

Exe 1.30

Escreva uma função em MATLAB/Octave cuja variável de entrada seja um número real do intervalo $]0, 1[$ em base decimal e cuja variável de saída seja a sua representação em base binária com uma mantissa com n dígitos (escolha um processo de arredondamento).

Obs 1.31

Como já se viu e já se disse, há números reais que não podem ser representados num sistema de vírgula flutuante normalizado, por maiores que sejam os parâmetros n, $|t_1|$ e $|t_2|$. Tem-se então a seguinte questão: como representar os números reais num sistema de vírgula flutuante normalizado cuja representação exata não existe? Um dos objetivos da norma IEEE 754, apresentada em 1985 pelo *Institute of Electrical and Electronics Engineers*, é responder a esta pergunta definindo regras de arredondamento. Outros aspetos que esta norma aborda são, entre outros, os formatos aritméticos e tratamento de exceções como por exemplo divisões por zero. Depois da definição de bit e byte, apresentam-se alguns elementos desta norma.

Def 1.32

(a) ⟦bit⟧ Chama-se bit (binary digit) a um segmento de memória capaz de armazenar o dígito binário zero ou o dígito binário um.

(b) ⟦byte⟧ Chama-se byte (contração de "by eight") a um segmento de memória capaz de armazenar oito bits.

Obs 1.33

Alguns elementos da norma IEEE 754:

(a) Representação binária.

(b) Precisão simples — representações com quatro bytes: 1 bit para o sinal, 23 bits para a mantissa e 8 bits para o expoente.

(c) Precisão dupla — representações com oito bytes: 1 bit para o sinal, 52 bits para a mantissa e 11 bits para o expoente.

(d) Tanto os maiores quanto os menores expoentes são reservados:

- Se todos os bits do expoente são uns: se a mantissa é 0 — representação do símbolo "infinito" (escrito em MATLAB como `Inf`) ; se mantissa não é 0 — representação do símbolo "não é um número" (escrito em MATLAB como `NaN`), que resulta de uma operação $0/0$, ∞/∞ ou $\infty - \infty$.
- Se todos os bits do expoente são zeros: a representação de precisão dupla muda de $1.f$ para $0.f$, permitindo alargar o número de representações.

Obs 1.33 (cont.)

(e) O maior número positivo de precisão dupla é `realmax` $= 1.7977 \times 10^{308}$ e o menor número positivo de precisão dupla é `realmin` $= 2.2251 \times 10^{-308}$.

(f) Outro número importante é o chamado de "epsilon da máquina" (escrito em MATLAB como `eps`) e é definido como a distância entre o número 1 e o número maior do que 1 e mais próximo de 1. Em precisão dupla, o epsilon da máquina é igual a `eps` $= 2.2204 \times 10^{-16}$.

Obs 1.34

Alguns comentários:

(a) O comprimento da mantissa e do expoente tem por objetivo conciliar dois objetivos conflituantes: poder representar números muito grandes e muito pequenos e que o espaçamento entre os números seja "pequeno".

(b) Note-se que o espaçamento entre números é uniforme entre potências de 2, mas muda por um fator de dois com cada potência adicional de dois. Por exemplo, o espaçamento de números entre um e dois é `eps`, entre dois e quatro é $2 \times$ `eps` e entre quatro e oito é $4 \times$ `eps`.

(c) Já se viu que nem todos os números reais podem ser representados exatamente em precisão dupla. Em muitas situações tal não é crítico, mas, é importante estar ciente de que um cálculo usando números inexatos que deveria resultar em um inteiro (como zero) pode não resultar devido a erros de arredondamento.

Exe 1.35

(a) Indique o espaço de memória de um vetor real de 10^6 elementos em precisão simples e em precisão dupla.

(b) Indique o espaço de memória de uma matriz real de ordem 10000 em precisão simples e em precisão dupla.

Res

(a) 4×10^6 bytes, ou 4MB, em precisão simples, e 8MB em precisão dupla.

(b) 4×10^8 bytes $= 0.4$GB em precisão simples, ou 0.8GB em precisão dupla.

Obs 1.36

Algumas das propriedades dos números reais deixam de ser válidas. Perde-se, por exemplo, a associatividade.

```
>> a = 1.0e308;
>> b = 1.1e308;
>> c = -1.001e308;
>> a + (b + c)
ans = 1.0990e+308
>> (a + b) + c
ans = Inf
```


Obs 1.37

O sistema não contém nem números arbitrariamente grandes, nem números arbitrariamente pequenos.

```
>> 10^(-10)
ans = 1.0000e-10
>> 10^(-1000)
ans = 0
```


Obs 1.38

Note-se que também há erros de arredondamento mesmo para números que nem são arbitrariamente grandes, nem números arbitrariamente pequenos.

```
>> 1/3
ans = 0.33333
```

Obs 1.39

Note-se, o entanto, que o resultado da operação 1/3 tem uma precisão maior do que a exibida acima. Para exibir um resultado com mais casas decimais, usa-se o comando `format`:

```
>> format long
>> 1/3
ans = 0.333333333333333
>> format short
>> 1/3
ans = 0.33333
```


Obs 1.40

Note-se que `format short` é a opção por defeito do MATLAB. O comando `help format` mostra as opções disponíveis. Por outro lado, o comando `sprintf` permite um maior controle sobre o número de casas decimais. Por exemplo, se se quiser exibir 1/3 com 30 caracteres (incluindo o ponto e o sinal), dos quais 20 são casas decimais da parte fracionária, usa-se:

```
>> sprintf('%30.20f', 1/3)
ans = 0.33333333333333331483
```


Obs 1.41

Note-se que no resultado abaixo alguns dígitos da representação do número 1/3 no MATLAB estão incorretos. Como já se viu, há números que possuem uma representação decimal finita mas cuja representação binária introduz erros de arredondamento.

```
>> sprintf('%30.20f', 0.125)
ans = 0.12500000000000000000
>> sprintf('%30.20f', 0.1)
ans = 0.10000000000000000555
```

O número 0.125 foi representado corretamente, enquanto o número 0.1 não. A falha na representação do número 0.1 se deve ao fato deste número ser uma dízima periódica na base binária, conforme (1).

Exe 1.42

Stéphane Clain, Maria T. Malheiro, GJM, Ricardo Costa, *Compact structural schemes for Ordinary Differential Equations* (em preparação) — resultados obtidos em precisão quádrupla e na linguagem Julia.

Tabela: Benchmark ODE1: $K = 2$.

R	N	ϕ		$\phi^{(1)}$		$\phi^{(2)}$	
		E	O	E	O	E	O
1	60	3.94E−11	—	3.94E−11	—	3.94E−11	—
	120	2.46E−12	4.0	2.46E−12	4.0	2.46E−12	4.0
	240	1.54E−13	4.0	1.54E−13	4.0	1.54E−13	4.0
5	60	2.12E−29	—	2.12E−29	—	2.12E−29	—
	120	5.18E−33	12.0	5.18E−33	12.0	5.18E−33	12.0
	240	1.26E−36	12.0	1.26E−36	12.0	1.26E−36	12.0

Def 1.43

⟦tipos de rros⟧

(a) Erros de medida (associado a problemas práticos).

(b) Erros de truncatura:

  (a) substituição de um problema contínuo por um problema discreto.

  (b) substituição de um processo de cálculo infinito por um finito.

(c) Erros de arredondamento — resultantes da representação de números reais em vírgula flutuantes e do uso de aritmética com precisão finita.

Def 1.44

Represente-se por x o valor exato de um número real e por x^* um seu valor aproximado. Então:

(a) erro de aproximação: $\Delta x^* = x - x^*$.

(b) erro absoluto: $|\Delta x^*| = |x - x^*|$.

(c) um limite superior (majorante) do erro absoluto: $\varepsilon \geqslant |\Delta x^*|$, tendo-se que $x \in [x^* - \varepsilon, x^* + \varepsilon]$ e também se escrevendo $x = x^* \pm \varepsilon$.

(d) erro relativo: $\frac{|x-x^*|}{|x|}$ para $x \neq 0$, que por vezes se aproxima por $\frac{|x-x^*|}{|x^*|}$ para $x^* \neq 0$.

(e) um limite superior (majorante) do erro relativo: $\varepsilon' \geqslant \frac{|x-x^*|}{|x|}$ para $x \neq 0$, que por vezes se aproxima por $\varepsilon' \geqslant \frac{|x-x^*|}{|x^*|}$ para $x^* \neq 0$.

Exe 1.45

Calcule os erros absolutos e os erros relativos nas seguintes situações:

(a) $x = -100$ e $x^* = -100.1$.

(b) $y = -1000000$ e $y^* = -1000000.1$.

Res

(a)

$$\text{erro absoluto: } |x - x^*| = 0.1$$

$$\text{erro relativo: } \frac{|x - x^*|}{|x|} = \frac{0.1}{|-100|} = 1 \times 10^{-3}$$

(b)

$$\text{erro absoluto: } |y - y^*| = 0.1$$

$$\text{erro relativo: } \frac{|y - y^*|}{|y|} = \frac{0.1}{|-1000000|} = 1 \times 10^{-7}$$

Obs 1.46

Seja f uma função real de variável real. Sendo x^* uma aproximação de x, que erros se cometem quando se calcula $f(x^*)$ em vez de $f(x)$? O seguinte teorema apresenta uma resposta, que se generaliza para funções reais de duas variáveis reais no teorema que se lhe segue.

Teo 1.47

Seja $f \in C^1(\mathbb{R};\mathbb{R})$. Seja, ainda, $x \in I_x = [x^* - \varepsilon_x, x^* + \varepsilon_x]$, em que x^* representa um valor aproximado do valor exato x, sendo ε_x um limite superior do erro absoluto. Então, quando se calcula $y^* = f(x^*)$ em vez de $y = f(x)$, tem-se

$$\varepsilon_y \leqslant \varepsilon_x M_x, \text{ com } M_x \geqslant \max_{x \in I_x} \left|f'(x)\right|.$$

Tem-se, ainda,

$$\varepsilon'_y \leqslant \frac{\varepsilon_x M_x}{|y|}.$$

Teo 1.48

Seja $f \in C^1(\mathbb{R}^2;\mathbb{R})$. Sejam, ainda, $x \in I_x = [x^* - \varepsilon_x, x^* + \varepsilon_x]$ e $y \in I_y = [y^* - \varepsilon_y, y^* + \varepsilon_y]$, em que x^* e y^* representam valores aproximados dos valores exatos x e y, respetivamente, sendo ε_x e ε_y limites superiores do erro absoluto. Então, quando se calcula $z^* = f(x^*, y^*)$ em vez de $z = f(x, y)$, tem-se

$$\varepsilon_z \leqslant \varepsilon_x M_x + \varepsilon_y M_y$$

com

$$M_x \geqslant \max_{x\in I_x, y\in I_y} \left|f'_x(x,y)\right| \text{ e } M_y \geqslant \max_{x\in I_x, y\in I_y} \left|f'_y(x,y)\right|.$$

Tem-se, ainda,

$$\varepsilon'_z \leqslant \frac{\varepsilon_x M_x + \varepsilon_y M_y}{|z|}.$$

Exe 1.49

Calcule um limite superior do erro absoluto e do erro relativo no cálculo da expressão $f(x) = \exp(x)$, sabendo que são usados os seguintes valores aproximados $x_1^* = 3.14$ com $\varepsilon_{x_1} = 0.005$ e $x_2^* = 3.1416$ com $\varepsilon_{x_2} = 0.00005$ de $x = \pi$.

Exe 1.50

Calcule um limite superior do erro absoluto e do erro relativo no cálculo da expressão $f(x,y) = \frac{x}{x+y}$, sabendo que são usados os seguintes valores aproximados $x^* = 3.14$ com $\varepsilon_x = 0.005$ de $x = \pi$ e $y^* = 1.732$ com $\varepsilon_y = 0.0005$ de $y = \sqrt{3}$.

Obs 1.51

Os teoremas anteriores generalizam-se da maneira natural quando f é uma função de mais do que duas variáveis reais.

Def 1.52

(a) ⟦estabilidade matemática⟧ Um problema diz-se matematicamente estável (ou bem condicionado) se pequenas perturbações nos dados provocarem pequenas perturbações nos resultados.

(b) ⟦estabilidade numérica⟧ Um algoritmo/método numérico diz-se numericamente estável para um problema bem condicionado se os erros de arredondamento cometidos no desenrolar do algoritmo/método numérico se propagarem de uma maneira controlada.

Obs 1.53

A estabilidade matemática é uma característica do problema e a estabilidade numérica é característica do algoritmo/método numérico.

Exe 1.54

Comente a seguinte informação na base no conceito da "estabilidade matemática":
Dada uma função real de variável real φ definamos o conjunto

$$S_\varphi = \{x^* \in \mathbb{R} : \varphi(x^*) = 0\}.$$

Então:

(i) se $f(x) = x^2 - 3x + 2$, tem-se $S_f = \{1, 2\}$.

(ii) se $g(x) = x^2 - 3.0001x + 2$, tem-se $S_g = \{0.9999, 2.0002\}$.

Exe 1.55

Comente a seguinte informação na base no conceito da "estabilidade matemática":

(i) A solução da equação diferencial $u''_{tt} + u''_{xx} = 0$, $t > 0$, $x \in \mathbb{R}$, com as condições

$$u(x,0) = 0 \quad \text{e} \quad u'_t(x,0) = 0, x \in \mathbb{R}$$

é $u(x,t) = 0$.

(ii) A solução da equação diferencial $\tilde{u}''_{tt} + \tilde{u}''_{xx} = 0$, $t > 0$, $x \in \mathbb{R}$, com as condições

$$\tilde{u}(x,0) = 0 \quad \text{e} \quad \tilde{u}'_t(x,0) = 10^{-4}\,\text{sen}(10^4 x), x \in \mathbb{R}$$

é $\tilde{u}(x,t) = 10^{-8}\,\text{sen}(10^4 x)\,\text{senh}(10^4 t)$.

Exe 1.56

Considere o sistema linear:

$$\begin{cases} 10^{-20}x_1 + x_2 = 1 \\ x_1 + x_2 = 2. \end{cases}$$

Comente a seguinte tabela, na base no conceito da "estabilidade numérica".

solução exata	solução pelo Método de Gauss	solução pelo Método de Gauss com escolha parcial de pivô
$x_1 = \dfrac{1}{1-10^{-20}}$	$x_1 = 0$	$x_1 = 1$
$x_2 = 1 - \dfrac{10^{-20}}{1-10^{-20}}$	$x_2 = 1$	$x_2 = 1$

Def 2.1

⟦zero de uma função real de variável real⟧ Seja f uma função real de variável real. Diz-se que x^* é um zero de f se $f(x^*) = 0$ (se f for um polinómio, é habitual chamar raízes aos zeros).

Def 2.2

Seja f uma função real de variável real.

(a) ⟦zero simples de uma função real de variável real⟧ Se

$$f(x^*) = 0 \quad \wedge \quad f'(x^*) \neq 0,$$

diz-se que x^* é um zero simples de f.

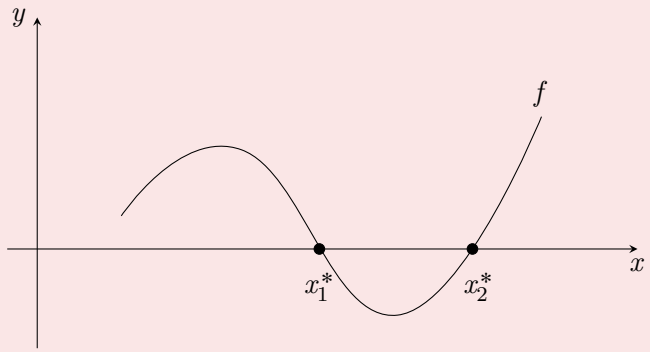
(b) ⟦zero com multiplicidade k de uma função real de variável real⟧ Se existe $k \in \mathbb{N}$ tal que

$$f(x^*) = f'(x^*) = \cdots = f^{(k-1)}(x^*) = 0 \quad \wedge \quad f^{(k)}(x^*) \neq 0,$$

diz-se que x^* é um zero múltiplo com multiplicidade k de f.

Obs 2.3

(a) Ilustração do conceito de zero de uma função real de variável real:



(b) Como todas as funções que se vão considerar são funções reais, passa-se apenas a dizer "Seja f uma função."

(c) Antes de estudar três métodos numéricos para determinar zeros de funções reais de variável real, recorda-se quatro teoremas relevantes para este capítulo.

Teo 2.4

[Teorema de Bolzano] Seja f uma função contínua em $[a, b]$. Se $f(a)f(b) < 0$, então f admite pelo menos um zero em $]a, b[$.

Teo 2.5

[Corolário do Teorema de Rolle] Seja f uma função contínua em $[a, b]$ e diferenciável em $]a, b[$. Se $f'(x) \neq 0$, para todo o $x \in]a, b[$, então f admite no máximo um zero em $]a, b[$.

Teo 2.6

[Teorema do valor médio de Lagrange] Seja f uma função contínua em $[a, b]$ e diferenciável em $]a, b[$. Então, existe $\xi \in]a, b[$ tal que

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

Def 2.7

Os métodos para a determinação de zeros de funções podem dividir-se em dois grupos.

(a) Métodos diretos: o(s) zero(s) da função é(são) determinado(s) de uma forma exata após um número finito de operações (supondo a utilização de aritmética exata). Um exemplo é a fórmula resolvente para polinómios de grau dois.

(b) Métodos iterativos: caracterizam-se por gerarem sucessões convergentes para o(s) zero(s) da função, distinguindo-se entre si pela forma como são geradas as sucessões de soluções aproximadas. O seu âmbito de aplicação é mais vasto que os métodos diretos.

Obs 2.8

Métodos iterativos:

(a) Seja x^* a solução do problema em causa.

(b) Regra geral, a solução exata é obtida através de um número infinito de iterações.

(c) O método iterativo é caracterizado por uma equação iterativa

$$x_{k+1} = \psi(x_k), \qquad k = 0, 1, \ldots.$$

(d) É necessária uma ou mais (depende do método) aproximações iniciais a x^* para iniciar o processo iterativo.

(e) O processo iterativo diz-se convergente se $\lim_{k\to\infty} x_k = x^*$.

(f) A convergência, em função dos valores iniciais, diz-se:

- local, se apenas se garantir convergência quando as aproximações iniciais estão suficientemente próximas de x^*.
- global, se se garantir convergência qualquer que seja a aproximação inicial pertencente ao domínio da função.

Obs 2.9

(g) É necessário um critério de paragem (CP), sendo condições naturais:

- proximidade entre duas aproximações consecutivas, na versão relativa ou absoluta

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leqslant \varepsilon_1 \quad \text{ou} \quad |x_{k+1} - x_k| \leqslant \varepsilon_2.$$

- a definição do problema dá origem a uma condição natural, que para a determinação de um zero de uma função é

$$|f(x_{k+1})| \leqslant \varepsilon_3.$$

- para prevenir situações de divergência ou de convergência lenta, é habitual impor um limite máximo no número de iterações, $k_{\max}$, especificado no início do processo iterativo.

Obs 2.10

(h) A convergência, quanto à rapidez, diz-se de ordem p se

$$\lim_{k\to\infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} = C \qquad (C \neq 0),$$

(a C chama-se constante de convergência assintótica).

Obs 2.11

Ideia do método das bisseções sucessivas: seja f uma função contínua em $[a, b]$ tal que $f(a)$ e $f(b)$ têm sinais diferentes, ou seja, $f(a)f(b) < 0$. O teorema de Bolzano permite afirmar que existe pelo menos um zero de f em $]a, b[$. Partindo-se de um intervalo com um único zero, a ideia é então ir dividindo sucessivamente os intervalos, escolhendo-se para a iteração seguinte o intervalo que contínua a conter o zero.

Alg 2.12 — Método das bisseções sucessivas

Input: $f \in C([a, b]; \mathbb{R})$, CP, $k_{\max} \in \mathbb{N}$
Output: $x_* \in \mathbb{R}$ ou "não convergiu"

```
1  a_0 ← a, b_0 ← b;
2  for k ← 0 to k_max − 1 do
3      x_{k+1} ← (a_k + b_k)/2;
4      if f(x_{k+1}) = 0 ∨ CP = V then
5          x_* ← x_{k+1};
6          return x_*;
7      else
8          if f(a_k)f(x_{k+1}) < 0 then
9              [a_{k+1}, b_{k+1}] ← [a_k, x_{k+1}];
10         else
11             [a_{k+1}, b_{k+1}] ← [x_{k+1}, b_k];
12         k ← k + 1;
13 return "não convergiu";
```

Obs 2.13

Note-se que x_* é uma aproximação de x^*, o zero de f em $[a, b]$.

Obs 2.14

(a) Uma alternativa à condição $f(x_{k+1}) = 0$ no algoritmo anterior é a condição $|f(x_{k+1})| \leqslant \varepsilon$, em que ε deve ser um valor adequado ao problema em questão.

(b) Um dos conceitos da definição seguinte vai ser relevante para a determinação do número de iterações do método das bisseções sucessivas quanso se especifica o erro absoluto máximo que se pretende obter.

Def 2.15

Seja $x \in \mathbb{R}$. Então:

(a) Chama-se *floor* de x, que se representa por $\lfloor x \rfloor$, a

$$\lfloor x \rfloor \stackrel{\text{def}}{=} n \in \mathbb{Z}, \text{ tal que } n \leqslant x \wedge n + 1 > x.$$

(b) Chama-se *ceil* de x, que se representa por $\lceil x \rceil$, a

$$\lceil x \rceil \stackrel{\text{def}}{=} n \in \mathbb{Z}, \text{ tal que } n \geqslant x \wedge n - 1 < x.$$

Exe 2.16

Calcule $\lfloor 2 \rfloor$, $\lfloor 2.3 \rfloor$, $\lfloor 2.7 \rfloor$, $\lceil 2 \rceil$, $\lceil 2.3 \rceil$ e $\lceil 2.7 \rceil$.

Res

$\lfloor 2 \rfloor = 2$, $\lfloor 2.3 \rfloor = 2$, $\lfloor 2.7 \rfloor = 2$, $\lceil 2 \rceil = 2$, $\lceil 2.3 \rceil = 3$ e $\lceil 2.7 \rceil = 3$.

Teo 2.17

Seja f uma função contínua em $[a, b]$ tal que $f(a)f(b) \leqslant 0$ e seja x^* o único zero de f em $[a, b]$. Então:

(a) A sucessão (x_k) gerada pelo método das bisseções sucessivas converge para x^*.

(b) O valor $\varepsilon_{k+1} = \frac{b-a}{2^{k+1}}$ constitui um majorante do erro absoluto de x_{k+1}.

Dem

(a)
- A sucessão (a_k) é monótona não decrescente e limitada por b, pelo que é convergente. Seja

$$\lim_{k\to\infty} a_k = \bar{a} \in [a, b].$$

- A sucessão (b_k) é monótona não crescente e limitada por a, pelo que é convergente. Seja

$$\lim_{k\to\infty} b_k = \bar{b} \in [a, b].$$

Dem (cont.)

- Como

$$\lim_{k\to\infty} (b_k - a_k) = \bar{b} - \bar{a}$$

e como

$$\lim_{k\to\infty} (b_k - a_k) = \lim_{k\to\infty} \frac{b-a}{2^k} = 0,$$

então $\bar{b} - \bar{a} = 0$, ou seja, $\bar{a} = \bar{b}$.
- Seja $\bar{x}$ o limite comum anterior. Como $x_{k+1} = \frac{a_k+b_k}{2}$, a sucessão (x_k) também é convergente com $\lim_{k\to\infty} x_k = \bar{x}$.
- Como $f(a_k)f(b_k) \leqslant 0$, para todo k, e f é contínua em $[a, b]$, então

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

- Assim, tem-se que $(f(\bar{x}))^2 \leqslant 0$.

Dem (cont.)

- Como a desigualdade anterior é válida se e só se $f(\bar{x}) = 0$, conclui-se que a sucessão (x_k) converge para x^*, o único zero f em $[a, b]$.

(b) • Uma vez que $x^* \in [a_k, b_k]$ e $x_{k+1} = \frac{a_k+b_k}{2}$, então

$$|x^* - x_{k+1}| \leqslant \frac{b_k - a_k}{2} = \frac{b-a}{2^{k+1}},$$

pelo que $\varepsilon_{k+1} = \frac{b-a}{2^{k+1}}$ é um majorante do erro absoluto de x_{k+1}.

Obs 2.18

Seja $\delta \in \mathbb{R}^+$ o erro absoluto máximo que se pretende obter. Então, como

$$\frac{b-a}{2^k} \leqslant \delta \Leftrightarrow k \geqslant \log_2\left(\frac{b-a}{\delta}\right)$$

é suficiente fazer $\left\lceil \log_2\left(\frac{b-a}{\delta}\right)\right\rceil$ iterações do método das bisseções.

Exe 2.19

Considere a função real de variável real $f(x) = 1 + x + \exp(x)$.

(a) Mostre que f tem um único zero e que esse zero pertence ao intervalo $[-2, -1]$.

(b) Indique o número mínimo de iterações do método das bisseções sucessivas por forma a garantir uma aproximação com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ do (único) zero de f.

(c) Recorrendo ao método das bisseções sucessivas, determine uma aproximação com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ do (único) zero de f (apresente os cálculos das iterações com 4 casas decimais).

Res

(a) Atendendo a que $f'(x) = 1 + \exp(x) > 0$ para $x \in \mathbb{R}$, tem-se que f é uma função monótona (crescente). Como $f(-2) = -0.865 < 0$ e $f(-1) = 0.368 > 0$, conclui-se, pelo teorema de Bolzano e pelo corolário do teorema de Rolle, que f tem um único zero e que esse zero pertence ao intervalo $[-2, -1]$.

(b) O número mínimo de iterações do método das bisseções sucessivas por forma a garantir uma aproximação com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ é

$$k = \left\lceil \log_2\left(\frac{b-a}{\delta}\right)\right\rceil = \left\lceil \log_2\left(\frac{-1-(-2)}{5 \times 10^{-3}}\right)\right\rceil = \lceil 7.6439 \rceil = 8.$$

Res (cont.)

(c) • Iteração 1 ($k = 0$): considerando $a_0 = -2$ e $b_0 = -1$, vem:

$$x_1 = \frac{-2 + (-1)}{2} = -1.5 \text{ e } f(-1.5) = -0.2769.$$

Como $f(-1.5)f(-2) > 0$, vem

$$[a_1, b_1] = [x_1, b_0] = [-1.5, -1].$$

• Iteração 2 ($k = 1$): considerando $a_1 = -1.5$ e $b_1 = -1$, vem:

$$x_2 = \frac{-1.5 + (-1)}{2} = -1.25 \text{ e } f(-1.25) = +0.0365.$$

Como $f(-1.25)f(-1.5) < 0$, vem

$$[a_2, b_2] = [a_1, x_2] = [-1.5, -1.25].$$

Res (cont.)

- Todas as iterações

k	a_k	$f(a_k)$	b_k	$f(b_k)$	$x_{k+1} = \frac{a_k+b_k}{2}$	$f(x_{k+1})$
0	−2.0000	−0.8647	−1.0000	0.3679	−1.5000	−0.2769
1	−1.5000	−0.2769	−1.0000	0.3679	−1.2500	0.0365
2	−1.5000	−0.2769	−1.2500	0.0365	−1.3750	−0.1222
3	−1.3750	−0.1222	−1.2500	0.0365	−1.3125	−0.0434
4	−1.3125	−0.0434	−1.2500	0.0365	−1.2812	−0.0036
5	−1.2812	−0.0036	−1.2500	0.0365	−1.2656	0.0164
6	−1.2812	−0.0036	−1.2656	0.0164	−1.2734	0.0064
7	−1.2812	−0.0036	−1.2734	0.0064	−1.2773	0.0014

Solução: $x^* \approx x_8 = -1.2773$.

Exe 2.20

Considere a função real de variável real $f(x) = x\ln(x) - 1$.

(a) Mostre que f tem um e só um zero e que esse zero pertence ao intervalo $[1, \mathrm{e}]$.

(b) Indique o número mínimo de iterações do método das bisseções sucessivas por forma a garantir uma aproximação com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ do (único) zero de f.

(c) Recorrendo ao método das bisseções sucessivas, determine uma aproximação com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ do (único) zero de f (apresente os cálculos das iterações com 5 casas decimais).

Exe 2.21

Considere a função real de variável real $f(x) = x + \ln(x)$.

(a) Mostre que f tem um e só um zero e que esse zero pertence ao intervalo $[0.5, 1.0]$.

(b) Recorrendo ao método das bisseções sucessivas, determine uma aproximação do (único) zero de f considerando para critério de paragem a condição

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leqslant 5 \times 10^{-3} \quad \wedge \quad |f(x_{k+1})| \leqslant 5 \times 10^{-4}$$

(apresente os cálculos das iterações com 5 casas decimais).

Exe 2.22

Implemente em MATLAB/Octave o método das bisseções sucessivas, considerando os seguintes critérios de paragem:

(a) v0: $k_{\max}$.

(b) v1: $k_{\max}$ e um majorante do erro absoluto δ.

(c) v2: $k_{\max}$ e $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$.

(d) v3: $k_{\max}$ e $|x_{k+1} - x_k| \leqslant \varepsilon_2$.

(e) v4: $k_{\max}$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(f) v5: $k_{\max}$, $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(g) v6: $k_{\max}$, $|x_{k+1} - x_k| \leqslant \varepsilon_2$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

Obs 2.23

Ideia do método iterativo simples:

(a) Para determinar zeros de funções através do método iterativo simples, também designado por iteração de ponto fixo, começa-se por reescrever o problema $f(x) = 0$ na forma $x = \varphi(x)$.

(b) Em seguida, escolhe-se um valor inicial x_0 e gera-se a sucessão (x_k) por intermédio da equação iteativa $x_{k+1} = \varphi(x_k)$, $k = 0, 1, \ldots$, razão pelo qual à função φ se chama função iteradora.

(c) A justificação do funcionamento deste método reside no seguinte argumento. Se a sucessão (x_k) convergir, para um dado valor x^*, e se a função iteradora φ for contínua, verifica-se então que $x^* = \varphi(x^*)$, ou seja, que x^* é um ponto fixo da função φ. Uma vez que, por hipótese, se tem que $f(x) = 0 \Leftrightarrow x = \varphi(x)$, conclui-se finalmente que $f(x^*) = 0$, ou seja, que o método iterativo simples, quando convergente, produz sucessões que convergem para zeros da função f.

Alg 2.24 — Método iterativo simples

Input: $\varphi \in C(\mathbb{R};\mathbb{R})$, tal que $f(x) = 0 \Leftrightarrow x = \varphi(x)$, $x_0 \in \mathbb{R}$, CP, $k_{\max} \in \mathbb{N}$
Output: $x_* \in \mathbb{R}$ ou "não convergiu"

```
for k ← 0 to k_max − 1 do
    x_{k+1} ← φ(x_k);
    if CP=V then
        x_* ← x_{k+1};
        return x_*;
    else
        k ← k + 1;
return "não convergiu";
```

Obs 2.25

Note-se que x_* é uma aproximação de x^*, o zero de f em $[a, b]$.

Teo 2.26

Seja φ uma função contínua em $[a, b]$ e diferenciável em $]a, b[$ tal que $L = \max_{x\in[a,b]} |\varphi'(x)| < 1$ e seja x^* o único zero de f em $[a, b]$. Então:

(a) Para qualquer valor inicial $x_0 \in [a, b]$, a sucessão (x_k) gerada pelo método iterativo simples converge para x^*.

(b) O valor $\varepsilon_{k+1} = \frac{L}{1-L}|x_{k+1} - x_k|$ constitui um majorante do erro absoluto de x_{k+1}.

Obs 2.27

A alínea (a) deste teorema indica que se a função iteradora for tal que $|\varphi'(x^*)| < 1$, o método iterativo simples converge desde que o valor inicial x_0 esteja suficientemente próximo de x^*. Das muitas (infinitas!) possibilidades de escolha de φ é necessário selecionar uma que verifique $|\varphi'(x)| < 1$ numa vizinhança de x^*.

Dem

(a)
- Seja $L = \max_{x\in[a,b]} |\varphi'(x)|$. Por hipótese, $L < 1$.
- Seja $x_0 \in [a, b]$.
- Como $x^* = \varphi(x^*)$ e $x_1 = \varphi(x_0)$, então $x_1 - x^* = \varphi(x_0) - \varphi(x^*)$.
- Assim, pelo teorema do valor médio de Lagrange, existe $\xi_0 \in [a, b]$ tal que

$$x_1 - x^* = \varphi'(\xi_0)(x_0 - x^*).$$

- Como $x_2 = \varphi(x_1)$, então $x_2 - x^* = \varphi(x_1) - \varphi(x^*)$, pelo que existe $\xi_1 \in [a, b]$ tal que que

$$x_2 - x^* = \varphi(x_1) - \varphi(x^*) = \varphi'(\xi_1)(x_1 - x^*)$$

pelo que

$$x_2 - x^* = \varphi'(\xi_1)\varphi'(\xi_0)(x_0 - x^*).$$

Dem (cont.)

- Continuando este raciocínio, existem $\xi_0, \ldots, \xi_{k-1} \in [a, b]$ tais que

$$x_k - x^* = \varphi'(\xi_{k-1})\varphi'(\xi_{k-2}) \cdots \varphi'(\xi_0)(x_0 - x^*).$$

- Tem-se, então, que

$$\begin{aligned} |x_k - x^*| &= |\varphi'(\xi_{k-1})||\varphi'(\xi_{k-2})| \cdots |\varphi'(\xi_0)||x_0 - x^*| \\ &\leqslant L^k |x_0 - x^*|. \end{aligned}$$

- Como $0 \leqslant L < 1$, então $\lim_{k\to\infty} L^k = 0$, pelo que $\lim_{k\to\infty} |x_k - x^*| = 0$, ou seja, a sucessão (x_k) converge para x^* para qualquer valor inicial $x_0 \in [a, b]$.

Dem (cont.)

(b)
- Pelo teorema do valor médio de Lagrange para a função φ no intervalo de extremos x_k e x^*, existe $\xi_k \in [\min(x_k, x^*), \max(x_k, x^*)]$ tal que

$$\varphi(x_k) - \varphi(x^*) = \varphi'(\xi_k)(x_k - x^*).$$

- Como $\varphi(x_k) = x_{k+1}$ e $\varphi(x^*) = x^*$, tem-se:

$$\begin{aligned} |x_{k+1} - x^*| &= |\varphi'(\xi_k)||x_k - x^*| \\ &= |\varphi'(\xi_k)||x_{k+1} - x^* + x_k - x_{k+1}|, \end{aligned}$$

pelo que, pela desigualdade triangular,

$$\begin{aligned} |x_{k+1} - x^*| &\leqslant L|x_{k+1} - x^* + x_k - x_{k+1}| \\ &\leqslant L\left(|x_{k+1} - x^*| + |x_k - x_{k+1}|\right). \end{aligned}$$

Dem (cont.)

- Tem-se, então, que

$$(1 - L)|x_{k+1} - x^*| \leqslant L|x_k - x_{k+1}|$$

ou seja,

$$|x_{k+1} - x^*| \leqslant \frac{L}{1 - L}|x_k - x_{k+1}|.$$

pelo que $\varepsilon_{k+1} = \frac{L}{1-L}|x_k - x_{k+1}|$ é um majorante do erro absoluto de x_{k+1}.

Exe 2.28

Considere a função $f(x) = 1 + x + \exp(x)$, que se sabe ter um único zero e que este pertence ao intervalo $[-2, -1]$.

(a) Determine uma função iteradora que torne o método iterativo simples convergente para qualquer $x_0 \in [-2, -1]$.

(b) Considerando a função iteradora da alínea anterior, aplique o método iterativo simples para determinar uma aproximação, com um erro absoluto inferior a $\delta = 5 \times 10^{-5}$, do (único) zero de f (apresente os cálculos das iterações com 5 casas decimais).

Res

(a) Por exemplo, $f(x) = 0 \Leftrightarrow x = -1 - \exp(x)$, Assim, fazendo $\varphi(x) = -1 - \exp(x)$ tem-se que $\varphi'(x) = -\exp(x)$, pelo que

$$L = \max_{x\in[-2,-1]} |\varphi'(x)| = \max_{x\in[-2,-1]} \exp(x) = \exp(-1) = 0.36788 < 1.$$

Assim, a função iteradora φ torna o método iterativo simples convergente para qualquer $x_0 \in [-2, -1]$.

Res (cont.)

(b) • Estimativa do erro

$$\varepsilon_{k+1} = \frac{L}{1-L}|x_{k+1} - x_k| = \frac{0.36788}{1-0.36788}|x_{k+1} - x_k| = 0.58198|x_{k+1} - x_k|.$$

- Critério de paragem: $\varepsilon_{k+1} \leqslant 5 \times 10^{-5}$.
- Iteração 1 ($k = 0$) fazendo, por exemplo, $x_0 = -1.5$:

$$x_1 = \varphi(x_0) = -1 - \exp(x_0) = -1.1.22313$$
$$\varepsilon_1 = 0.58198|x_1 - x_0| = 1.6 \times 10^{-1} > 5 \times 10^{-5}.$$

- Iteração 2 ($k = 1$):

$$x_2 = \varphi(x_1) = -1 - \exp(x_1) = -1.29431$$
$$\varepsilon_2 = 0.58198|x_2 - x_1| = 4.1 \times 10^{-2} > 5 \times 10^{-5}.$$

Res (cont.)

- Todas as iterações

k	x_k	$x_{k+1} = -1 - \exp(x_k)$	$\varepsilon_{k+1} = 0.58198\lvert x_{k+1} - x_k\rvert$
0	-1.50000	-1.22313	$1.6\times10^{-1} > \delta$
1	-1.22313	-1.29431	$4.1\times10^{-2} > \delta$
2	-1.29431	-1.27409	$1.2\times10^{-2} > \delta$
3	-1.27409	-1.27969	$3.3\times10^{-3} > \delta$
4	-1.27969	-1.27812	$9.1\times10^{-4} > \delta$
5	-1.27812	-1.27856	$2.5\times10^{-4} > \delta$
6	-1.27856	-1.27844	$7.0\times10^{-5} > \delta$
7	-1.27844	-1.27847	$2.0\times10^{-5} \leqslant \delta$

- Solução: $x^* \approx x_8 = -1.27847$.

Exe 2.29

Considere a função $f(x) = \exp(x) - 3x$.

(a) Mostre que f tem dois zeros.

(b) Determine uma função iteradora e uma aproximação inicial que torne o método iterativo simples convergente para um dos zeros.

(c) Considerando a função iteradora da alínea anterior, aplique o método iterativo simples para determinar uma aproximação a esse zero com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ (apresente os cálculos das iterações com 5 casas decimais).

(d) Determine uma função iteradora e uma aproximação inicial que torne o método iterativo simples convergente para o outro zero.

(e) Considerando a função iteradora da alínea anterior, aplique o método iterativo simples para determinar uma aproximação a esse zero com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ (apresente os cálculos das iterações com 5 casas decimais).

Exe 2.30

Considere a função $f(x) = \exp(x)\ln(x) - 1$.

(a) Mostre que f tem um único zero.

(b) Determine uma função iteradora e uma aproximação inicial que torne o método iterativo simples convergente para o zero de f.

(c) Considerando a função iteradora da alínea anterior, aplique o método iterativo simples para determinar uma aproximação ao zero de f com um erro absoluto inferior a $\delta = 5 \times 10^{-3}$ (apresente os cálculos das iterações com 5 casas decimais).

Exe 2.31

Considere a função $f(x) = x + \ln(x)$.

(a) Mostre que f tem um único zero.

(b) Pretende-se determinar o seu zero recorrendo ao método iterativo simples. Para tal, considere as seguintes funções:

$$\varphi_1(x) = -\ln(x),$$
$$\varphi_2(x) = \exp(-x),$$
$$\varphi_3(x) = \frac{x + \exp(-x)}{2}.$$

(i) Mostre que φ_1, φ_2 e φ_3 são funções iteradoras de f.

(ii) Indique, justificando, quais das funções iteradoras tornam o método iterativo simples convergente.

(iii) Indique, justificando, qual das funções iteradoras deve ser usada para calcular o zero de f.

Exe 2.31 (cont.)

(iv) Aplique o método iterativo simples com a função iteradora identificada na alínea anterior e considerando para critério de paragem a condição

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leqslant 1 \times 10^{-2} \quad \wedge \quad |f(x_{k+1})| \leqslant 5 \times 10^{-4}$$

(apresente os cálculos das iterações com 5 casas decimais).

Exe 2.32

Implemente em MATLAB/Octave o método iterativo simples, considerando os seguintes critérios de paragem:

(a) v0: $k_{\max}$.

(b) v1: $k_{\max}$ e um majorante do erro absoluto δ dado L.

(c) v2: $k_{\max}$ e $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$.

(d) v3: $k_{\max}$ e $|x_{k+1} - x_k| \leqslant \varepsilon_2$.

(e) v4: $k_{\max}$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(f) v5: $k_{\max}$, $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(g) v6: $k_{\max}$, $|x_{k+1} - x_k| \leqslant \varepsilon_2$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

Obs 2.33

Ideia do método de Newton:

(a) O método de Newton é um dos métodos mais poderosos para resolver equações do tipo $f(x) = 0$. Tal como no caso do método iterativo simples (de que pode ser considerado um caso particular), este método parte de uma estimativa inicial x_0 e gera uma sucessão (x_k) de uma forma recorrente.

(b) O novo valor da sucessão, x_{k+1}, é determinado como sendo a abcissa do ponto de intersecção com o eixo dos xx da recta tangente ao gráfico da função no ponto $(x_k, (f(x_k))$, ou seja, no ponto correspondente ao valor anterior da sucessão.

(c) A expressão de recorrência do método de Newton vem então dada por

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

Alg 2.34 — Método de Newton

Input: $f \in C^1([a,b];\mathbb{R})$, $x_0 \in [a,b]$, CP, $k_{\max} \in \mathbb{N}$
Output: $x_* \in \mathbb{R}$ ou "não convergiu"

for $k \leftarrow 0$ **to** $k_{\max} - 1$ **do**
  $x_{k+1} \leftarrow x_k - \frac{f(x_k)}{f'(x_k)}$;
  if *CP=V* **then**
    $x_* \leftarrow x_{k+1}$;
    return x_*;
  else
    $k \leftarrow k+1$;
return *"não convergiu"*;

Obs 2.35

Note-se que x_* é uma aproximação de x^*, o zero de f em $[a,b]$.

Teo 2.36

[Teorema de Taylor (que é uma extensão do Teorema do valor médio de Lagrange)] Seja $f \in C^n([a,b];\mathbb{R})$ e $f^{(n+1)}$ definida em $]a,b[$. Seja, ainda, $c \in [a,b]$. Então, existe $\xi \in]a,b[$ tal que

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}.$$

Obs 2.37

(a) Teorema de Taylor para $n = 0$:

$$f(x) = f(c) + f'(\xi)(x-c).$$

(b) Teorema de Taylor para $n = 1$:

$$f(x) = f(c) + f'(c)(x-c) + \frac{f''(\xi)}{2}(x-c)^2.$$

Teo 2.38

Seja $f \in C^2([a,b];\mathbb{R})$ tal que x^* é o único zero de f em $[a,b]$.

(a) Se
  (i) $\forall x \in [a,b]\ [f'(x) \neq 0]$,
  (ii) $\forall x \in [a,b]\ [f''(x) \leqslant 0 \vee f''(x) \geqslant 0]$, e
  (iii) $\exists x_0 \in [a,b]\ [f(x_0)f''(x_0) \geqslant 0]$,
  então a sucessão gerada pelo método de Newton com aproximação inicial x_0 converge monotonamente para x^*.

(b) Seja (x_k) uma sucessão gerada pelo método de Newton convergente para x^* com $x_k \in [a,b]$, $k = 0, 1, \ldots$. Sejam, ainda,

$$M_2 = \max_{x\in[a,b]} |f''(x)| \text{ e } m_1 = \min_{x\in[a,b]} |f'(x)|.$$

Então, se $m_1 > 0$, o valor $\varepsilon_{k+1} = \frac{M_2}{2m_1}|x_{k+1} - x_k|^2$ constitui um majorante do erro absoluto de x_{k+1}.

Dem

(a)
- Comece-se por considerar o caso (i.1) $f'(x) > 0$, ou seja, f crescente em $[a,b]$, e (ii.2) $f''(x) \geqslant 0$ (nos outros casos a demonstração é semelhante).
- Seja $x_0 \in [a,b]$. Então, $f(x_0) \geqslant 0$ por (ii.2) e (iii), pelo que $x^* \leqslant x_0$ por (i.1).
- Como $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$, então $x_1 \leqslant x_0$ pois $\frac{f(x_0)}{f'(x_0)} > 0$.
- Pelo teorema de Taylor, existe $\xi_0 \in]a,b[$ tal que

$$f(x^*) = f(x_0) + f'(x_0)(x^* - x_0) + \frac{f''(\xi_0)}{2}(x^* - x_0)^2.$$

- Como $f(x^*) = 0$, então

$$x^* - x_0 = -\frac{f(x_0)}{f'(x_0)} - \frac{f''(\xi_0)}{2f'(x_0)}(x^* - x_0)^2.$$

Dem

- Como $\frac{f''(\xi_0)}{2f'(x_0)}(x^* - x_0)^2 > 0$ por (i.1) e (ii,2), tem-se

$$x^* - x_0 \leqslant -\frac{f(x_0)}{f'(x_0)} \Leftrightarrow x^* \leqslant x_0 - \frac{f(x_0)}{f'(x_0)} \Leftrightarrow x^* \leqslant x_1,$$

  pelo que também $f(x_1) \geqslant 0$ por (i.1).
- Resumindo, tem-se:

$$x^* \leqslant x_1 \leqslant x_0 \quad \text{e} \quad f(x_1) \geqslant 0.$$

- Admita-se, agora, que $x^* \leqslant x_k$ e que $f(x_k) \geqslant 0$. Então, repetindo os argumentos anteriores, tem-se que

$$x^* \leqslant x_{k+1} \leqslant x_k \quad \text{e} \quad f(x_{k+1}) \geqslant 0.$$

- Assim, (x_k) é uma sucessão decrescente e limitada inferiormente por x^*, pelo que é convergente.

Dem (cont.)

- Seja

$$\lim_{k\to\infty} x_k = \bar{x} \in [a, b].$$

- Como $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ e f e f' são funções contínuas, então no limite tem-se $\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}$.
- Então, $f(\bar{x}) = 0$. Como f tem um único zero em $[a, b]$, conclui-se que $\bar{x} = x^*$, ou seja, que a sucessão gerada pelo método de Newton converge monotonamente para x^*.

Dem (cont.)

(b)
- Como $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$, então

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0 \qquad (\square).$$

- Pelo teorema de Taylor, existe $\xi_k \in]a, b[$ tal que

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

- Substituindo $(\square)$ em $(\square\square)$, tem-se que

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

Dem (cont.)

- Pelo teorema de Taylor, existe $\zeta_k \in]a, b[$ tal que

$$f(x_{k+1}) = f(x^*) + f'(\zeta_k)(x_{k+1} - x^*),$$

  pelo que, atendendo a que $f(x^*) = 0$,

$$f(x_{k+1}) = f'(\zeta_k)(x_{k+1} - x^*) \qquad (**).$$

- Combinando as expressões (*) e (**), tem-se

$$|f'(\zeta_k)||x_{k+1} - x^*| = \frac{|f''(\xi_k)|}{2}|x_{k+1} - x_k|^2,$$

  pelo que

$$|x_{k+1} - x^*| = \frac{|f''(\xi_k)|}{2|f'(\zeta_k)|}|x_{k+1} - x_k|^2.$$

Dem (cont.)

- Como $M_2 = \max_{x\in[a,b]} |f''(x)|$ e $m_1 = \min_{x\in[a,b]} |f'(x)| > 0$, então

$$|x_{k+1} - x^*| \leqslant \frac{M_2}{2m_1}|x_{k+1} - x_k|^2,$$

pelo que $\varepsilon_{k+1} = \frac{M_2}{2m_1}|x_{k+1} - x_k|^2$ constitui um majorante do erro absoluto de x_{k+1}.

Exe 2.39

Considere a função $f(x) = 1 + x + \exp(x)$, que se sabe ter um único zero e que este pertence ao intervalo $[-2, -1]$.

(a) Determine $x_0 \in [-2, -1]$ tal que as condições suficientes de convergência do método de Newton se verifiquem.

(b) Aplique o método de Newton considerando a aproximação inicial identificada na alínea anterior para determinar uma aproximação ao zero de f com um erro absoluto inferior a $\delta = 5\times10^{-6}$ (apresente os cálculos das iterações com 5 casas decimais).

Res

(a)
- Como $f'(x) = 1 + \exp(x)$, então, $f'(x) \neq 0$ para $x \in [-2, -1]$.
- Como $f''(x) = \exp(x)$, então, $f''(x) \geqslant 0$ para $x \in [-2, -1]$.
- Para $f(x_0)f''(x_0) \geqslant 0$, tem que se ter $f(x_0) \geqslant 0$ pelo que, por exemplo, pode-se considerar $x_0 = -1$ pois $f(-1) = 0.36788$.

Res (cont.)

(b)
- Estimativa do erro: atendendo a

$$m_1 = \min_{x\in[-2,-1]} |f'(x)| = 1 + \exp(-2) = 1.13534$$
$$M_2 = \max_{x\in[-2,-1]} |f''(x)| = \exp(-1) = 0.36788$$
$$\frac{M_2}{2m_1} = 0.16201,$$

tem-se que $\varepsilon_{k+1} = 0.16201|x_{k+1} - x_k|^2$ é um majorante do erro de x_{k+1}.
- Critério de paragem: $\varepsilon_{k+1} \leqslant 5 \times 10^{-6}$.
- Iteração 1 ($k = 0$)

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -1.26894$$
$$\varepsilon_1 = 0.16201|x_1 - x_0|^2 = 1.2 \times 10^{-1} > \delta.$$

Res (cont.)

- Iteração 2 ($k = 1$)

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -1.27485$$
$$\varepsilon_2 = 0.16201|x_2 - x_1|^2 = 1.5 \times 10^{-5} > \delta.$$

- Iteração 3 ($k = 2$)

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = -1.27486$$
$$\varepsilon_3 = 0.16201|x_3 - x_2|^2 = 1.6 \times 10^{-11} \leqslant \delta.$$

- Solução: $x^* \approx x_3 = -1.27846$.

Exe 2.40

Considere a função $f(x) = x\ln(x) - 1$, que se sabe ter um único zero e que este pertence ao intervalo $[1, \mathrm{e}]$.

(a) Determine $x_0 \in [1, \mathrm{e}]$ tal que as condições suficientes de convergência do método de Newton se verifiquem.

(b) Aplique o método de Newton considerando a aproximação inicial identificada na alínea anterior para determinar uma aproximação ao zero de f com um erro absoluto inferior a $\delta = 5\times10^{-3}$ (apresente os cálculos das iterações com 5 casas decimais).

Exe 2.41

Considere a função $f(x) = x + \ln(x)$, que se sabe ter um único zero e que este pertence ao intervalo $[0.5, 1.0]$.

(a) Determine $x_0 \in [0.5, 1.0]$ tal que as condições suficientes de convergência do método de Newton se verifiquem.

(b) Aplique o método de Newton considerando a aproximação inicial identificada na alínea anterior para determinar uma aproximação ao zero de f considerando para critério de paragem a condição

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leqslant 5 \times 10^{-3} \quad \wedge \quad |f(x_{k+1})| \leqslant 5 \times 10^{-4}$$

(apresente os cálculos das iterações com 5 casas decimais).

Exe 2.42

Considere a função $f(x) = x - \exp(-x)$, que se sabe ter um único zero e que este pertence ao intervalo $[0, 1]$.

(a) Determine $x_0 \in [0, 1]$ tal que as condições suficientes de convergência do método de Newton se verifiquem.

(b) Aplique o método de Newton considerando a aproximação inicial identificada na alínea anterior para determinar uma aproximação ao zero de f considerando para critério de paragem a condição

$$|x_{k+1} - x_k| \leqslant 5 \times 10^{-4}$$

(apresente os cálculos das iterações com 5 casas decimais).

Exe 2.43

Considere a função $f(x) = 0.51x - \mathrm{sen}(x)$.

(a) Faça três iterações do método de Newton considerando $x_0 = 1$ (apresente os cálculos das iterações com 5 casas decimais).

(b) Faça três iterações do método de Newton considerando $x_0 = 2$ (apresente os cálculos das iterações com 5 casas decimais).

(c) Comente os resultados obtidos nas duas alíneas anteriores.

Exe 2.44

Implemente em MATLAB/Octave o método de Newton, considerando os seguintes critérios de paragem:

(a) v0: $k_{\max}$.

(b) v1: $k_{\max}$ e um majorante do erro absoluto δ dados M_2 e m_1.

(c) v2: $k_{\max}$ e $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$.

(d) v3: $k_{\max}$ e $|x_{k+1}-x_k| \leqslant \varepsilon_2$.

(e) v4: $k_{\max}$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(f) v5: $k_{\max}$, $\frac{|x_{k+1}-x_k|}{|x_{k+1}|} \leqslant \varepsilon_1$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

(g) v6: $k_{\max}$, $|x_{k+1}-x_k| \leqslant \varepsilon_2$ e $|f(x_{k+1})| \leqslant \varepsilon_3$.

Obs 2.45

E como se comportam os métodos apresentados relativamente à rapidez da sua convergência? Recorde-se que um método diz-se de ordem p se

$$\lim_{k\to\infty} \frac{|x^*-x_{k+1}|}{|x^*-x_k|^p} = C.$$

Então, a partir de certo k deve-se ter

$$\frac{|x^*-x_{k+1}|}{|x^*-x_k|^p} \approx C \quad \text{e} \quad \frac{|x^*-x_k|}{|x^*-x_{k-1}|^p} \approx C,$$

pelo que

Obs 2.45 (cont.)

$$\begin{aligned}
&\frac{|x^*-x_{k+1}|}{|x^*-x_k|^p} \approx \frac{|x^*-x_k|}{|x^*-x_{k-1}|^p} \\
&\Leftrightarrow |x^*-x_{k+1}||x^*-x_{k-1}|^p \approx |x^*-x_k||x^*-x_k|^p \\
&\Leftrightarrow \ln\left(|x^*-x_{k+1}||x^*-x_{k-1}|^p\right) \approx \ln\left(|x^*-x_k||x^*-x_k|^p\right) \\
&\Leftrightarrow \ln|x^*-x_{k+1}| + p\ln|x^*-x_{k-1}| \approx \ln|x^*-x_k| + p\ln|x^*-x_k| \\
&\Leftrightarrow \ln|x^*-x_{k+1}| - \ln|x^*-x_k| \approx p\ln|x^*-x_k| - p\ln|x^*-x_{k-1}| \\
&\Leftrightarrow \ln\frac{|x^*-x_{k+1}|}{|x^*-x_k|} \approx p\ln\frac{|x^*-x_k|}{|x^*-x_{k-1}|} \\
&\Leftrightarrow p \approx \frac{\ln\frac{|x^*-x_{k+1}|}{|x^*-x_k|}}{\ln\frac{|x^*-x_k|}{|x^*-x_{k-1}|}} = \frac{\ln\frac{\Delta_{k+1}}{\Delta_k}}{\ln\frac{\Delta_k}{\Delta_{k-1}}}.
\end{aligned}$$

Exe 2.46

Sabe-se que a única raiz real do polinómio $p(x) = x^3 - 2x^2 + x + 4$ é $x^* = -1$.

(a) Faça quatro iterações do método iteratio simples considerando a função iteradora $\varphi(x) = -0.1x^3+0.2x^2+0.9x-0.4$ e a aproximação inicial $x_0 = -1.25$ e estime a ordem de convergência.

(b) Faça quatro iterações do método de Newton considerando a aproximação inicial $x_0 = -1.25$ e estime a ordem de convergência.

Res

	(a)			(b)	
k	Δ_k	p	k	Δ_k	p
0	2.50E−01	—	0	2.50E−01	—
1	1.72E−02	0.62	1	3.22E−02	1.92
2	3.29E−03	0.98	2	6.29E−04	1.99
3	6.52E−04	1.00	3	2.47E−07	2.00
4	1.30E−04	—	4	3.82E−14	—

Exe 2.47

Sabendo que $x^* = 1$ é uma raiz de multiplicidade dois do polinómio $p(x) = (x-1)^2$, aplique cinco iterações do método de Newton considerando a aproximação inicial $x_0 = 0.1$ e estime a ordem de convergência.

Res

k	Δ_k	p
0	9.00E−01	—
1	4.50E−01	1.00
2	2.25E−01	1.00
3	1.13E−01	1.00
4	5.63E−02	—

Teo 2.48

(a) Método iterativo simples: convergência linear ou de 1ª ordem para zeros simples

$$\Delta_{k+1} \propto \Delta_k.$$

(b) Método de Newton: convergência quadrática ou de 2ª ordem para zeros simples

$$\Delta_{k+1} \propto \Delta_k^2.$$

(c) Método de Newton: convergência linear ou de 1ª ordem para zeros com multiplicidade superior a um

$$\Delta_{k+1} \propto \Delta_k.$$

Obs 3.1

Neste capítulo vai-se começar por recordar vários conceitos introduzidos no capítulo de "Sistemas de equações lineares" da unidade curricular de "Álgebra Linear". Depois vai-se avançar no estudo de sistemas de equações lineares tendo em consideração o uso de aritmética com precisão finita.

Def 3.2

⟦equação linear, incógnitas ou variáveis, termo independente ou segundo membro⟧ Uma equação linear nas incógnitas ou variáveis $x_1, x_2, \ldots, x_n \in \mathbb{R}$ é uma equação do tipo

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b,$$

onde $a_1, a_2, \ldots, a_n, b \in \mathbb{R}$. A b chama-se termo independente ou segundo membro da equação linear.

Obs 3.3

A equação linear

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

nas incógnitas $x_1, x_2, \ldots, x_n$ pode ser escrita na forma matricial

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b \end{bmatrix}.$$

Def 3.4

⟦sistema de equações lineares⟧ A um conjunto finito de equações lineares chama-se sistema de equações lineares (ou simplesmente sistema ou sistema linear, caso não resulte ambíguo).

Exe 3.5

Dê um exemplo de um sistema com duas equações lineares e com três incógnitas.

Res

$$\begin{cases} x + 2y + z = 1 \\ 3x - y + z = 0. \end{cases}$$

Def 3.6

⟦matriz dos coeficientes, vetor dos termos independentes, vetor das incógnitas, matriz aumentada ou matriz ampliada, conjunto solução⟧ Seja (S) o sistema de m equações lineares nas n incógnitas $x_1, x_2, \ldots, x_n$ dado por

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n = b_2 \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \quad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \cdots + a_{mn}x_n = b_m. \end{cases}$$

Então:

(a) à matriz $A = [a_{ij}] \in \mathcal{M}_{m\times n}(\mathbb{R})$ chama-se matriz dos coeficientes de (S).

(b) à matriz coluna $b = [b_i] \in \mathcal{M}_{m\times 1}(\mathbb{R})$ chama-se vetor dos termos independentes de (S).

Def 3.6 (cont.)

(c) à matriz coluna $x = [x_i] \in \mathcal{M}_{n\times 1}(\mathbb{R})$ chama-se vetor das incógnitas de (S).

(d) à matriz

$$A|b \overset{\text{def}}{=} \left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} & b_m \end{array}\right].$$

chama-se matriz aumentada ou matriz ampliada de (S).

(e) Chama-se conjunto solução do sistema (S), que se representa por $\mathrm{CS}_{(S)}$, a

$$\mathrm{CS}_{(S)} \overset{\text{def}}{=} \{(x_1, \ldots, x_n) \in \mathbb{R}^n : A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = b\}.$$

Obs 3.7

Note-se que o sistema (S) da definição anterior pode ser escrito na forma matricial

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix},$$

ou, em notação matricial, como $Ax = b$.

Def 3.8

Seja (S) um sistema de equações lineares.

(a) ⟦sistema possível ou sistema compatível ou sistema consistente⟧ Diz-se que (S) é um sistema possível (que se abrevia por "Pos") ou sistema compatível ou sistema consistente se $\#\,\mathrm{CS}_{(S)} > 0$.

(b) ⟦sistema possível e determinado⟧ Diz-se que (S) é um sistema possível e determinado (que se abrevia por "PD") se $\#\,\mathrm{CS}_{(S)} = 1$.

(c) ⟦sistema possível e indeterminado⟧ Diz-se que (S) é um sistema possível e indeterminado (que se abrevia por "PI") se $\#\,\mathrm{CS}_{(S)} = +\infty$.

(d) ⟦sistema impossível ou sistema incompatível ou sistema inconsistente⟧ Diz-se que (S) é um sistema impossível (que se abrevia por "IMP") ou sistema incompatível ou sistema inconsistente se $\#\,\mathrm{CS}_{(S)} = 0$.

Teo 3.9

Um sistema de equações lineares ou é PD ou PI ou IMP.

Def 3.10

⟦Sistema de Cramer⟧ Um sistema de equações lineares diz-se um sistema de Cramer se o número de equações é igual ao número de incógnitas e é um sistema PD.

Teo 3.11

Seja (S) um sistema de equações lineares cuja matriz dos coeficientes é uma matriz quadrada A. Então, (S) é um sistema de Cramer se e só se $\det(A) \neq 0$.

Obs 3.12

(a) Neste curso só se vão considerar sistemas de Cramer.

(b) Se $Ax = b$ é um sistemas de Cramer, então A é uma matriz invertível e $x = A^{-1}b$. No entanto "nunca" se resolve um sistema de equações lineares através deste processo pois há métodos computacionalmente menos exigentes.

(c) Métodos a estudar — métodos diretos: métodos em que se obtém a solução (exata) do sistema após um número finito de operações (considerando que as operações são com aritmética exata) — método de Gauss (MG) e método de Gauss com escolha parcial de pivô (MGPP).

(d) Os métodos diretos que se vão estudar, não sendo os únicos, formam a base de um conjunto alargado de outros métodos que tiram partido da estrutura da matriz dos coeficientes. Há também uma vasta literatura de métodos iterativos para a resolução de sistemas de equações lineares, mas que não vão ser abordados neste curso.

Obs 3.13

Determinar o conjunto solução de um sistema de Cramer cuja matriz dos coeficientes é uma matriz triangular superior ou inferior é simples.

(a) Se é uma matriz triangular superior: começa-se por determinar o valor da última incógnita através da última equação; depois, determina-se o valor da penúltima incógnita substituindo-se na penúltima equação o valor da última incógnita, repetindo-se o processo até se determinar o valor da primeira incógnita. A este algoritmo chama-se "Método de substituição de trás para a frente" — MeSTaF.

(b) Se é uma matriz triangular inferior, determinar o seu conjunto solução é simples: começa-se por determinar o valor da primeira incógnita através da primeira equação; depois, determina-se o valor da segunda incógnita substituindo-se na segunda equação o valor da primeira incógnita, repetindo-se o processo até se determinar o valor da última incógnita. A este algoritmo chama-se "Método de substituição da frente para trás" — MeSFeT.

Exe 3.14

Determine o conjunto solução do sistema de Cramer (S) dado por

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

Res

Como a matriz dos coeficientes do sistema (S) é triangular superior, aplicando-se o MeSTaF, tem-se:

- $2x_3 = 4 \Leftrightarrow x_3 = 2$;
- $3x_2 - 4x_3 = 4 \Leftrightarrow 3x_2 - 4 \times (2) = 4 \Leftrightarrow x_2 = 4$;
- $x_1 - 2x_2 + 3x_3 = -1 \Leftrightarrow x_1 - 2 \times (4) + 3 \times (2) = -1 \Leftrightarrow x_1 = 1$,

pelo que $\mathrm{CS}_{(S)} = \{(1, 4, 2)\}$.

Exe 3.15

Determine o conjunto solução do sistema de Cramer (S)

$$\begin{cases} 2.21x_1 + 3.04x_2 - 1.27x_3 = 0.03 \\ 3.21x_2 - 2.78x_3 = 1.73 \\ 4.23x_3 = -6.41. \end{cases}$$

Obs 3.16

Seja A a matriz dos coeficientes de um sistema de Cramer. Se A não é uma matriz triangular superior, os métodos que se consideram neste curso aplicam um conjunto de operações, ditas "operações elementares" tais que o sistema resultante tem o mesmo conjunto solução e a matriz dos coeficientes transformou-se numa matriz triangular superior. Recorda-se de seguida estas operações.

Def 3.17

⟦operação elementar do tipo I nas linhas de uma matriz⟧ Sejam $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ e $i, i' \in \{1, \ldots, m\}$ tal que $i \neq i'$. Chama-se operação elementar do tipo I nas linhas da matriz A à troca de duas linhas. A troca de ℓ_i com $\ell_{i'}$ representa-se por $\ell_i \leftrightarrow \ell_{i'}$.

Exe 3.18

Indique a matriz que se obtém depois de aplicar a operação do tipo I $\ell_1 \leftrightarrow \ell_3$ à matriz $A = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & -1 & 1 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix}$.

Res

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

Def 3.19

⟦operação elementar do tipo II nas linhas de uma matriz⟧ Sejam $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, $i \in \{1,\dots,m\}$ e $\alpha \in \mathbb{R}-\{0\}$. Chama-se operação elementar do tipo II nas linhas da matriz A à substituição de uma linha por um seu múltiplo não-nulo. A substituição de ℓ_i pela linha que se obtém multiplicando por α os elementos de ℓ_i representa-se por $\ell_i \leftarrow \alpha\ell_i$, que se lê "$\ell_i$ toma valor de $\alpha\ell_i$".

Exe 3.20

Indique a matriz que se obtém depois de aplicar a operação do tipo II $\ell_3 \leftarrow \frac{1}{2}\ell_3$ à matriz $A = \left[\begin{smallmatrix} 1 & 2 & 0 & 1 \\ 0 & -1 & 1 & 1 \\ 2 & 2 & 1 & 0 \end{smallmatrix}\right]$.

Res

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

Def 3.21

⟦operação elementar do tipo III nas linhas de uma matriz⟧ Sejam $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, $i, i' \in \{1,\dots,m\}$ e $\beta \in \mathbb{R}$. Chama-se operação elementar do tipo III nas linhas da matriz A à substituição de uma linha pela sua soma com um múltiplo de outra linha. A substituição de ℓ_i pela linha que se obtém somando os elementos de ℓ_i aos elementos que se obtêm multiplicando por β os elementos de $\ell_{i'}$ representa-se por $\ell_i \leftarrow \ell_i + \beta\ell_{i'}$, que se lê "$\ell_i$ toma valor de $\ell_i + \beta\ell_{i'}$".

Exe 3.22

Indique a matriz que se obtém depois de aplicar a operação do tipo III $\ell_1 \leftarrow \ell_1 - \frac{1}{2}\ell_2$ à matriz $A = \left[\begin{smallmatrix} 0 & -1 & 1 & 1 \\ 2 & 2 & 1 & 0 \end{smallmatrix}\right]$.

Res

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

Def 3.23

⟦matrizes equivalentes⟧ Sejam $A, B \in \mathcal{M}_{m\times n}(\mathbb{R})$. Diz-se que A e B são matrizes equivalentes, escrevendo-se $A \longleftrightarrow B$, se se pode obter uma a partir da outra através duma sequência (finita) de operações elementares (com linhas).

Exe 3.24

Seja a matriz $A = \left[\begin{smallmatrix} 0 & 2 & 4 & 0 \\ 1 & 1 & 0 & 2 \\ 2 & 2 & 0 & 5 \end{smallmatrix}\right]$. Efetue a seguinte sequência de operações na matriz A: $\ell_1 \leftrightarrow \ell_2$, $\ell_3 \leftarrow \ell_3 - 2\ell_1$, $\ell_1 \leftarrow \ell_1 - 2\ell_3$, $\ell_2 \leftarrow \frac{1}{2}\ell_2$ e $\ell_1 \leftarrow \ell_1 - \ell_2$.

Res

$$\begin{bmatrix} 0 & 2 & 4 & 0 \\ 1 & 1 & 0 & 2 \\ 2 & 2 & 0 & 5 \end{bmatrix} \xleftrightarrow[\ell_1 \leftrightarrow \ell_2]{} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 4 & 0 \\ 2 & 2 & 0 & 5 \end{bmatrix} \xleftrightarrow[\ell_3 \leftarrow \ell_3 - 2\ell_1]{}$$

$$\begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xleftrightarrow{\ell_1 \leftarrow \ell_1 - 2\ell_3} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xleftrightarrow[\ell_2 \leftarrow \frac{1}{2}\ell_2]{}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xleftrightarrow{\ell_1 \leftarrow \ell_1 - \ell_2} \begin{bmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Def 3.25

⟦matriz elementar⟧ Seja $E \in \mathcal{M}_{n\times n}(\mathbb{R})$. Diz-se que E é uma matriz elementar se se pode obter através de uma operação elementar sobre a matriz I_n.

Exe 3.26

A partir de I_4, determine as matrizes elementares obtidas através das seguintes operações elementares:

(a) $\ell_2 \leftrightarrow \ell_4$.

(b) $\ell_3 \leftarrow 2\ell_3$.

(c) $\ell_3 \leftarrow \ell_3 - 2\ell_1$.

Res

(a) $\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} \xleftrightarrow[\ell_2 \leftrightarrow \ell_4]{} \begin{bmatrix}1&0&0&0\\0&0&0&1\\0&0&1&0\\0&1&0&0\end{bmatrix} = P.$

(b) $\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} \xleftrightarrow[\ell_3 \leftarrow 2\ell_3]{} \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&2&0\\0&0&0&1\end{bmatrix} = M.$

(c) $\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} \xleftrightarrow[\ell_3 \leftarrow \ell_3 - 2\ell_1]{} \begin{bmatrix}1&0&0&0\\0&1&0&0\\-2&0&1&0\\0&0&0&1\end{bmatrix} = L.$

Teo 3.27

Aplicar uma operação elementar a uma matriz corresponde a pré-multiplicar essa matriz pela matriz elementar correspondente a essa operação elementar.

Exe 3.28

Ilustre o teorema anterior considerando a matriz $A = \begin{bmatrix}1&2&0&-1\\2&2&-1&-1\\1&1&3&2\\2&1&1&-2\end{bmatrix}$ e as operações elementares do exercício Exe 3.26.

Res

(a)

- Processo 1: aplicar a operação $\ell_2 \leftrightarrow \ell_4$

$$\begin{bmatrix}1&2&0&-1\\2&2&-1&-1\\1&1&3&2\\2&1&1&-2\end{bmatrix} \xleftrightarrow[\ell_2 \leftrightarrow \ell_4]{} \begin{bmatrix}1&2&0&-1\\2&1&1&-2\\1&1&3&2\\2&2&-1&-1\end{bmatrix} = A_P.$$

- Processo 2: calcular PA

$$\begin{bmatrix}1&0&0&0\\0&0&0&1\\0&0&1&0\\0&1&0&0\end{bmatrix}\begin{bmatrix}1&2&0&-1\\2&2&-1&-1\\1&1&3&2\\2&1&1&-2\end{bmatrix} = \begin{bmatrix}1&2&0&-1\\2&1&1&-2\\1&1&3&2\\2&2&-1&-1\end{bmatrix} = A_P.$$

Res (cont.)

(b)
- Processo 1: aplicar a operação $\ell_3 \leftarrow 2\ell_3$

$$\begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ 1 & 1 & 3 & 2 \\ 2 & 1 & 1 & -2 \end{bmatrix} \xrightarrow[\ell_3 \leftarrow 2\ell_3]{} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ 2 & 2 & 6 & 4 \\ 2 & 1 & 1 & -2 \end{bmatrix} = A_M.$$

- Processo 2: calcular MA

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ 1 & 1 & 3 & 2 \\ 2 & 1 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ 2 & 2 & 6 & 4 \\ 2 & 1 & 1 & -2 \end{bmatrix} = A_M.$$

Res (cont.)

(c)
- Processo 1: aplicar a operação $\ell_3 \leftarrow \ell_3 - 2\ell_1$.

$$\begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ 1 & 1 & 3 & 2 \\ 2 & 1 & 1 & -2 \end{bmatrix} \xrightarrow[\ell_3 \leftarrow \ell_3 - 2\ell_1]{} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 2 & -1 & -1 \\ -1 & -3 & 3 & 4 \\ 2 & 1 & 1 & -2 \end{bmatrix} = A_L.$$

- Processo 2: calcular LA

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

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

Teo 3.29

(a) Seja P uma matriz elementar associada a uma operação elementar do tipo I. Então, P é invertível e $P^{-1} = P$.

(b) Seja M uma matriz elementar associada a uma operação elementar do tipo II $\ell_i \leftarrow \beta\ell_i$. Então, M é invertível e M^{-1} é a matriz identidade exceto na posição ii que vale $\frac{1}{\beta}$.

(c) Seja L uma matriz elementar associada operação elementar do tipo III $\ell_i \leftarrow \ell_i + \beta\ell_{i'}$. Então, L é invertível e L^{-1} é a matriz identidade exceto na posição ii' que vale β.

Def 3.30

[[sistemas de equações lineares equivalentes]] Dois sistemas de equações lineares dizem-se equivalentes se tiverem o mesmo conjunto solução.

Teo 3.31

Sejam (S) um sistema de equações lineares e (S') um sistema de equações lineares cuja matriz aumentada foi obtida a partir da matriz aumentada de (S) através de um cojunto finito de operações do tipo, I, II e III. Então, (S) e (S') são equivalentes.

Obs 3.32

O teorema anterior justifica o método de Gauss (algoritmo Alg. 3.35) na versão para sistemas de Cramer. Este método baseia-se no "Algoritmo Transformação em Escada" (ATEsc). Este algoritmo, quando aplicado a matrizes quadradas, através de operações elementares do tipo I e III, devolve uma matriz triangular superior.

Alg 3.33 — "Algoritmo Transformação em Escada" (ATEsc)

Input: matriz $A = [a_{ij}] \in \mathcal{M}_{m\times n}(\mathbb{R})$
Output: uma matriz em escada equivalente à matriz A

Passo 1 [inicializar o algoritmo]
$i \leftarrow 1$, $j \leftarrow$ índice da coluna não-nula mais à esquerda da matriz A

Passo 2 [selecionar o elemento pivô]
se $a_{ij} = 0$ **então**
$\ell_i \leftrightarrow \ell_k$, em que ℓ_k é a primeira linha abaixo da linha ℓ_i com um elemento diferente de zero na coluna c_j

Passo 3 [anular os elementos abaixo do pivô]
para $p \leftarrow i+1$ **até** m **fazer**
$m_{pj} \leftarrow \frac{a_{pj}}{a_{ij}}$, $\quad \ell_p \leftarrow \ell_p - m_{pj}\ell_i$

Passo 4 [terminar?]
se já se obteve uma matriz em escada **então** terminar
senão
$i \leftarrow i+1$, $j \leftarrow$ índice da coluna não-nula mais à esquerda da matriz $A(i:\text{end},:)$
ir para o Passo 2

Exe 3.34

Aplique o ATEsc à matriz

$$A = \begin{bmatrix} 1 & 1 & 0.5 & 1 \\ -2 & -2 & 0 & -1 \\ 1 & -1 & 3 & -1 \end{bmatrix}.$$

Res

$$\begin{bmatrix} 1 & 1 & 0.5 & 1 \\ -2 & -2 & 0 & -1 \\ 1 & -1 & 3 & -1 \end{bmatrix} \xleftrightarrow[\substack{\ell_2 \leftarrow \ell_2+2\ell_1 \\ \ell_3 \leftarrow \ell_3-\ell_1}]{} \begin{bmatrix} 1 & 1 & 0.5 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & -2 & 3.5 & -2 \end{bmatrix} \xleftrightarrow{\ell_2 \leftrightarrow \ell_3} \begin{bmatrix} 1 & 1 & 0.5 & 1 \\ 0 & -2 & 2.5 & -2 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Alg 3.35 — "Algoritmo do método de Gauss para sistemas de Cramer" (MG)

Input: matriz aumentada $A|b$ de um sistema de Cramer (S)
Output: $\text{CS}_{(S)}$

Passo 1 [ATEsc]
aplicar o ATEsc à matriz aumentada $A|b$

Passo 2 [determinar $\text{CS}_{(S)}$]
determinar o valor das incógnitas através da aplicação do MeSTaF à matriz obtida no Passo 1

Exe 3.36

Resolva através do MG o sistema de Cramer (S) dado por

$$\begin{cases} 1.1301x_1 - 2.0234x_2 + 2.9891x_3 = -1.2123 \\ 1.8734x_1 - 1.3412x_2 + 1.9561x_3 = 2.0345 \\ 3.1234x_1 + 0.8978x_2 + 2.0125x_3 = 2.7986. \end{cases}$$

Res

Passo 1 Aplicação do ATEsc:

$$\left[\begin{array}{rrr|r} 1.1301 & -2.0234 & 2.9891 & -1.2123 \\ 1.8734 & -1.3412 & 1.9561 & 2.0345 \\ 3.1234 & 0.8978 & 2.0125 & 2.7986 \end{array}\right] \longleftrightarrow \begin{array}{l} \\ \ell_2 \leftarrow \ell_2 - 1.6577\ell_1 \\ \ell_3 \leftarrow \ell_3 - 2.7638\ell_1 \end{array}$$

$$\left[\begin{array}{rrr|r} 1.1301 & -2.0234 & 2.9891 & -1.2123 \\ 0 & 2.0130 & -2.9990 & 4.0442 \\ 0 & 6.4901 & -6.2489 & 6.1492 \end{array}\right] \longleftrightarrow \begin{array}{l} \\ \\ \ell_3 \leftarrow \ell_3 - 3.2240\ell_2 \end{array}$$

$$\left[\begin{array}{rrr|r} 1.1301 & -2.0234 & 2.9891 & -1.2123 \\ 0 & 2.0130 & -2.9990 & 4.0442 \\ 0 & 0 & 3.4201 & -6.8893 \end{array}\right]$$

Passo 2 Aplicação do MeSTaF:

- $x_3 = -2.0144$.
- $x_2 = -0.9920$.
- $x_1 = 2.4791$,

pelo que $\mathrm{CS}_{(S)} = \{(2.4791, -0.9920, -2.0144)\}$.

Obs 3.37

Se se utiliza aritmética com precisão finita o MG pode apresentar problemas. O exemplo seguinte ilustra estas dificuldades.

Exe 3.38

Considere o sistema de Cramer

$$\begin{cases} 10^{-10}x_1 + x_2 = 1 \\ \quad x_1 + x_2 = 2. \end{cases}$$

(a) Mostre que $\mathrm{CS}_{(S)} = \left\{\left(\frac{1}{1-10^{-10}}, 1 - \frac{10^{-10}}{1-10^{-10}}\right)\right\}$.

(b) Resolva-o pelo MG.

Res

(a) Fazendo $x_1 = \frac{1}{1-10^{-10}}$ e $x_2 = 1 - \frac{10^{-10}}{1-10^{-10}}$

- na primeira equação tem-se $10^{-10} \times \frac{1}{1-10^{-10}} + 1 - \frac{10^{-10}}{1-10^{-10}} = 1$ e
- na segunda equação tem-se $\frac{1}{1-10^{-10}} + 1 - \frac{10^{-10}}{1-10^{-10}} = 2$.

(b) **Passo 1**

$$\left[\begin{array}{rr|r} 10^{-10} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right] \longleftrightarrow \begin{array}{l} \\ \ell_2 \leftarrow \ell_2 - 10^{10}\ell_1 \end{array} \left[\begin{array}{rr|r} 10^{-10} & 1 & 1 \\ 0 & -10^{-10} & -10^{-10} \end{array}\right]$$

Passo 2

Aplicando-se o MeSTaF, tem-se:

- $x_2 = 1$.
- $x_1 = 0$.

Obs 3.39

A razão do problema é a existência de pivôs, em valor absoluto, muito pequenos, o que implica multiplicadores muito grandes, o que pode implicar uma grande ampliação dos erros provocados pela aritmética com precisão finita. Um procedimento para tentar resolver estes problemas consiste em trocar linhas por forma a evitar pivôs com valores absolutos muito pequenos. Esta ideia dá origem ao "Algoritmo do método de Gauss com escolha parcial de pivô (MGPP) para sistemas de Cramer", em que a diferença relativa ao MG reside no Passo 2 do ATEsc.

Alg 3.40 — "Algoritmo Transformação em Escada — v2" (ATEsc-v2)

Input: matriz $A = [a_{ij}] \in \mathcal{M}_{m\times n}(\mathbb{R})$
Output: uma matriz em escada equivalente à matriz A

Passo 1 [inicializar o algoritmo]
$i \leftarrow 1$, $j \leftarrow$ índice da coluna não-nula mais à esquerda da matriz A

Passo 2 [selecionar o elemento pivô]
$k \leftarrow \underset{\bar{k}\in\{i,\ldots,m\}}{\arg\max} |a_{\bar{k}i}|$
se $i \neq k$ **então** $\ell_i \leftrightarrow \ell_k$

Passo 3 [anular os elementos abaixo do pivô]
para $p \leftarrow i+1$ **até** m **fazer**
$m_{pj} \leftarrow \frac{a_{pj}}{a_{ij}}$, $\quad \ell_p \leftarrow \ell_p - m_{pj}\ell_i$

Passo 4 [terminar?]
se já se obteve uma matriz em escada **então** terminar
senão
$i \leftarrow i+1$, $j \leftarrow$ índice da coluna não-nula mais à esquerda da matriz $A(i:\text{end},:)$
ir para o Passo 2

Alg 3.41 — "Algoritmo do método de Gauss com escolha parcial de pivô para sistemas de Cramer" (MGPP)

Input: matriz aumentada $A|b$ de um sistema de Cramer (S)
Output: $\text{CS}_{(S)}$

Passo 1 [ATEsc-v2]
aplicar o ATEsc-v2 à matriz aumentada $A|b$

Passo 2 [determinar $\text{CS}_{(S)}$]
determinar o valor das incógnitas através da aplicação do MeSTaF à matriz obtida no Passo 1

Exe 3.42

Resolva o seguinte sistema de Cramer através do MGPP:

$$\begin{cases} 4x_1 + 13x_2 + 2x_3 = -15 \\ -8x_1 + 10x_2 + 8x_3 = 6 \\ 2x_1 + 6.5x_2 + 5.5x_3 = -3. \end{cases}$$

Res

Passo 1 Aplicação do ATEsc-v2:

$$\left[\begin{array}{ccc|c} 4 & 13 & 2 & -15 \\ -8 & 10 & 8 & 6 \\ 2 & 6.5 & 5.5 & -3 \end{array}\right] \xleftrightarrow[\ell_1 \leftrightarrow \ell_2]{} \left[\begin{array}{ccc|c} -8 & 10 & 8 & 6 \\ 4 & 13 & 2 & -15 \\ 2 & 6.5 & 5.5 & -3 \end{array}\right]$$

$$\xleftrightarrow[\substack{\ell_2 \leftarrow \ell_2 + 0.5\ell_1 \\ \ell_3 \leftarrow \ell_3 + 0.25\ell_1}]{} \left[\begin{array}{ccc|c} -8 & 10 & 8 & 6 \\ 0 & 18 & 6 & -12 \\ 0 & 9 & 7.5 & -1.5 \end{array}\right] \xleftrightarrow[\ell_3 \leftarrow \ell_3 - 0.5\ell_2]{}$$

$$\left[\begin{array}{ccc|c} -8 & 10 & 8 & 6 \\ 0 & 18 & 6 & -12 \\ 0 & 0 & 4.5 & 4.5 \end{array}\right]$$

Passo 2 Aplicação do MeSTaF:

- $x_3 = 1.$
- $x_2 = -1.$
- $x_1 = -1,$

pelo que $\text{CS}_{(S)} = \{(-1,-1,1)\}$.

Exe 3.43

Resolva o seguinte sistema de Cramer através do MGPP:

$$\begin{cases} 4x_1 + 2x_3 = -15 \\ -8x_1 + 10x_2 + 8x_3 = 6 \\ 20x_1 - x_2 + x_3 = 0. \end{cases}$$

Exe 3.44

Resolva o seguinte sistema de Cramer através do MGPP:

$$\begin{cases} 3x_1 + 4x_2 + 7x_3 + 20x_4 = 504 \\ 20x_1 + 25x_2 + 40x_3 + 50x_4 = 1170 \\ 10x_1 + 15x_2 + 20x_3 + 22x_4 = 970 \\ 10x_1 + 8x_2 + 10x_3 + 15x_4 = 601. \end{cases}$$

Exe 3.45

Seja (S) o sistema de Cramer

$$\begin{cases} 10^{-6}x_1 + x_3 = 1 \\ x_1 + 10^{-6}x_2 + 2x_3 = 3 \\ x_1 + 2x_2 - x_3 = 2. \end{cases}$$

(a) Resolva (S) através do MG.

(b) Resolva (S) através do MGPP.

Exe 3.46

Seja $\varepsilon \in \mathbb{R} - \{0\}$. Seja, ainda, (S_ε) o sistema de Cramer

$$\begin{cases} \varepsilon x_1 + x_2 = 1 \\ x_1 + x_2 = 2. \end{cases}$$

(a) Mostre que $\mathrm{CS}_{(S)} = \left\{\left(\frac{1}{1-\varepsilon}, 1 - \frac{\varepsilon}{1-\varepsilon}\right)\right\}$.

(b) Resolva-o pelo MG, fazendo $\varepsilon = 10^{-20}$, considerando aritmética exata.

(c) Resolva-o pelo MG, fazendo $\varepsilon = 10^{-20}$, considerando aritmética com precisãofinia finita.

(d) Resolva-o por MGPP, fazendo novamente $\varepsilon = 10^{-20}$, considerando aritmética com precisão finita.

(e) Comente os resultados obtidos.

Exe 3.47

Calcule a inversa da matriz invertível $A = \begin{bmatrix} 2.1 & -1.2 & 4.3 \\ 6.1 & 3.2 & -7.3 \\ 4.8 & 1.7 & 3.3 \end{bmatrix}$ através do MGPP.

Exe 3.48

(a) Vetorize em MATLAB/Octave a linha "$\ell_p \leftarrow \ell_p - m_{pj}\ell_i$" dos algoritmos ATEsc e ATEsc-v2.

(b) Implemente em MATLAB/Octave o MG.

(c) Implemente em MATLAB/Octave o MGPP.

Obs 3.49

Em muitas situações é necessário resolver muitos sistemas de equações lineares com a mesma matriz dos coeficientes. Os diferentes vetores dos termos independentes podem ou não ser conhecidos *a priori*. Vai-se agora apresentaestudar um método para o caso negativo.

Def 3.50

⟦fatorização ou decomposição LU⟧ Seja A uma matriz quadrada. À fatorização ou decomposição $A = LU$, se existir, em que L é uma matriz triangular inferior e U é uma matriz triangular superior chama-se fatorização LU ou decomposição LU de A.

Obs 3.51

(a) O nome vem do inglês: "L" de "lower" e "U" de "upper".

(b) Nem todas as matrizes quadradas admitem uma fatorização LU. O seguinte teorema indica uma condição necessária e suficiente para existir a fatorização.

Teo 3.52

Seja $A = [a_{ij}] \in \mathcal{M}_{n\times n}(\mathbb{R})$. Então, existe uma fatorização LU de A se e só se os determinantes das matrizes

$$A_k = [a_{ij}] \in \mathcal{M}_{k\times k}(\mathbb{R}), k = 1, \ldots, n-1,$$

são todos diferentes de zero.

Obs 3.53

Seja A uma matriz de ordem n tal que admite uma fatorização LU.

(a) Então existe uma infinidade de fatorizações, pois existem $n(n+1)$ incógnitas e n^2 equações (independentes). Há, pois, n graus de liberdade.

(b) Uma das fatorizações LU que existe é a chamada "Fatorização LU de Doolittle", onde as n condições extra que se impõem são $(L)_{ii} = 1$, $i = 1, \ldots, n$.

Obs 3.53 (cont.)

(c) Vai-se agora mostrar que a fatorização LU de Doolittle está relacionada com o algoritmo ATEsc no caso de aí não haver troca de linhas, ou seja, só haver operações elementares do tipo III:

- Sejam L_k, $k = 1, \ldots, n-1$, os produtos das matrizes elementares que se realizaram no Passo 2. Então, L_k é a matriz identidade de ordem n exceto nas posições $(k+1, k), \ldots, (n, k)$ onde se tem os multiplicadores, ou seja,

$$(L_k)_{k+1,k} = -m_{k+1,k}, \ldots, (L_k)_{n,k} = -m_{n,k}.$$

- , Então, tem-se

$$(L_{n-1} \cdots L_1)A = U,$$

ou ainda

$$\bar{L}A = U, \quad \bar{L} = L_{n-1} \cdots L_1.$$

Obs 3.53 (cont.)

- Como as matrizes L_k são triangulares inferiores e como o produto de duas matrizes triangulares inferiores ainda é uma matriz triangular inferior, então $\bar{L}$ também é uma matriz triangular inferior.
- As matrizes L_k, $k = 1, \ldots, n-1$, são matrizes invertíveis, pelo que $\bar{L}$ também é uma matriz invertível. Assim, tem-se que

$$\bar{L}A = U \Leftrightarrow A = LU, \quad \text{com } L = \bar{L}^{-1}.$$

- Pode-se mostrar que L_k^{-1} é a matriz identidade de ordem n exceto nas posições $(k, k+1), \ldots, (k, n)$ onde se tem

$$(L_k^{-1})_{k,k+1} = m_{k,k+1}, \ldots, (L_k^{-1})_{k,n} = m_{k,n}.$$

Obs 3.53 (cont.)

- Seja $k \in \{1, \ldots, n-1\}$. Então, pode-se mostrar que L_k^{-1} é a matriz identidade de ordem n exceto nas posições $(k+1, k), \ldots, (n, k)$ onde se tem

$$(L_k^{-1})_{k+1,k} = m_{k+1,k}, \ldots, (L_k^{-1})_{n,k} = m_{n,k}.$$

- A matriz $L = [\ell_{ij}]\mathcal{M}_{n\times n}(\mathbb{R})$ é a matriz triangular inferior dada por

$$\ell_{ij} = \begin{cases} 1, & \text{se } i = j, \\ 0, & \text{se } i < j, \\ m_{ij}, & \text{se } i > j. \end{cases}$$

Exe 3.54

Seja a matriz $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$.

(a) Aplique o ATEsc à matriz A.

(b) Indique as matrizes das operações elementares que realizou bem como o produto das matrizes elementares em cada Passo 2 do ATEsc.

(c) Calcule a matriz L da fatorização LU de Doolittle.

(d) Verifique que $A = LU$.

Res

(a)

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{matrix} \longleftarrow \\ \ell_2 \leftarrow \ell_2 - \ell_1 \\ \ell_3 \leftarrow \ell_3 + \ell_1 \end{matrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -2 & -1 \\ 0 & 1 & 2 \end{bmatrix} \begin{matrix} \longleftarrow \\ \\ \ell_3 \leftarrow \ell_3 + 0.5\ell_2 \end{matrix}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -2 & -1 \\ 0 & 0 & 1.5 \end{bmatrix} = U.$$

(b) • $\ell_2 \leftarrow \ell_2 - \ell_1$, $\ell_3 \leftarrow \ell_3 + \ell_1$:

$$L_{11} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, L_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, L_1 = L_{11}L_{12} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

• $\ell_3 \leftarrow \ell_3 + 0.5\ell_2$:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

Res (cont.)

(c)

$$\begin{aligned} L &= (L_2L_1)^{-1} \\ &= L_1^{-1}L_2^{-1} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -0.5 & 1 \end{bmatrix}, \end{aligned}$$

ou seja, confirma-se que a matriz L é uma matriz triangular inferior cujos elementos da diagonal são uns e em que os elementos abaixo da diagonal são os multiplicadores do algoritmo ATEsc com os sinais trocados.

Res (cont.)

(d)

$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -0.5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -2 & -1 \\ 0 & 0 & 1.5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix} = A.$$

Exe 3.55

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

(a) Aplique o ATEsc à matriz A.

(b) Indique as matrizes das operações elementares que realizou.

(c) Calcule a matriz L da fatorização LU de Doolittle.

(d) Verifique que $A = LU$.

Obs 3.56

Está-se agora em condições de resolver sistemas de Cramer $Ax = b$ através de uma fatorização LU (em particular da fatorização LU de Doolittle) a tendendo a que $Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb$:

Passo 1 — Resolver o sistema triangular inferior $Ly = b$ para y através do algoritmo MeSFeT.

Passo 2 — Resolver o sistema triangular superior $Ux = y$ para x através do algoritmo MeSTaF.

Exe 3.57

Sejam a matriz invertível $A = \begin{bmatrix} 4 & 2 & 7 \\ 3 & 5 & -6 \\ 1 & -3 & 2 \end{bmatrix}$ e as matrizes coluna $b_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$ e $b_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$.

(a) Determine as matrizes L e U da fatorização LU de Doolittle.

(b) Resolva os sistemas de Cramer $Ax = b_1$ e $Ax = b_2$ recorrendo à fatorização LU de Doolittle.

Exe 3.58

Implemente em MATLAB/Octave a fatorização LU de Doolittle (se houver necessidade de troca de linhas, a função deve terminar e indicar uma mensagem de erro).

Obs 3.59

E se há operações elementares do tipo I nos algoritmos ATEsc ou ATEsc-v2? Analise-se o seguinte exemplo.

Exe 3.60

Seja a matriz $A = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix}$.

(a) Aplique o ATEsc-v2 à matriz A.

(b) Indique as matrizes das operações elementares que realizou.

Res

(a)

$$\begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix} \xleftrightarrow[\ell_1 \leftrightarrow \ell_2]{} \begin{bmatrix} 4 & 1 & 1 \\ 1 & 2 & 4 \\ 2 & 4 & 1 \end{bmatrix} \xleftrightarrow[\substack{\ell_2 \leftarrow \ell_2 - 0.25\ell_1 \\ \ell_3 \leftarrow \ell_3 - 0.5\ell_1}]{} \begin{bmatrix} 4 & 1 & 1 \\ 0 & 1.75 & 3.75 \\ 0 & 3.5 & 0.5 \end{bmatrix} \xleftrightarrow[\ell_2 \leftrightarrow \ell_3]{} \begin{bmatrix} 4 & 1 & 1 \\ 0 & 3.5 & 0.5 \\ 0 & 1.75 & 3.75 \end{bmatrix} \xleftrightarrow[\ell_3 \leftarrow \ell_3 - 0.5\ell_2]{} \begin{bmatrix} 4 & 1 & 1 \\ 0 & 3.5 & 0.5 \\ 0 & 0 & 3.5 \end{bmatrix}$$

Res (cont.)

(b)
- $\ell_1 \leftrightarrow \ell_2$:

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- $\ell_2 \leftarrow \ell_2 - 0.25\ell_1$, $\ell_3 \leftarrow \ell_3 - 0.5\ell_1$:

$$L_{11} = \begin{bmatrix} 1 & 0 & 0 \\ 0.25 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, L_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}, L_1 = L_{11}L_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0.25 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}.$$

- $\ell_2 \leftrightarrow \ell_3$:

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

- $\ell_3 \leftarrow \ell_3 - 0.5\ell_2$:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

Obs 3.61

Note-se que

$$\begin{aligned} & L_2P_2L_1P_1A = U \\ & \quad \Leftrightarrow L_2P_2L_1(P_2^{-1}P_2)P_1A = U \\ & \quad \Leftrightarrow L_2(P_2L_1P_2^{-1})(P_2P_1)A = U \\ & \quad \Leftrightarrow \bar{L}PA = U \end{aligned}$$

com

$$\begin{aligned} & P = P_2P_1, \\ & L_2' = L_2, L_1' = P_2L_1P_2^{-1}, \\ & \bar{L} = L_2'L_1'. \end{aligned}$$

Como $\bar{L}$ é uma matriz invertível, tem-se que

$$\bar{L}PA = U \Leftrightarrow PA = LU, \quad \text{com } L = \bar{L}^{-1}.$$

Obs 3.61 (cont.)

No caso geral, tem-se

$$PA = LU,$$

com

$$\begin{aligned} & P = P_{n-1}P_{n-2}\dots P_2P_1, \\ & L = \left(L_{n-1}'L_{n-2}'\dots L_2'L_1'\right)^{-1}, \\ & L_k' = P_{n-1}\dots P_{k+1}L_kP_{k+1}^{-1}\dots P_{n+1}^{-1} \\ & \quad = P_{n-1}\dots P_{k+1}L_kP_{k+1}\dots P_{n+1}, k = 1, \dots, n-1. \end{aligned}$$

Def 3.62

⟦fatorização ou decomposição PLU⟧ Seja A uma matriz quadrada. À fatorização ou decomposição $PA = LU$ em que P é uma matriz de permutação (ou seja, a matriz identidade eventualmente com linhas trocadas), L é uma matriz triangular inferior e U é uma matriz triangular superior chama-se fatorização PLU ou decomposição PLU de A.

Obs 3.63

(a) O nome vem do inglês: "P" de "permutation", "L" de "lower" e "U" de "upper".

(b) Pode-se motrar que todas as matrizes quadradas admitem uma fatorização PLU.

(c) Uma das fatorizações PLU que existe é a chamada "Fatorização PLU de Doolittle", onde os elementos da diagonal de L são uns.

(d) A fatorização apresentada na observação Obs 3.obs:PLU é a fatorização PLU de Doolittle.

Exe 3.64

Seja a matriz $A = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix}$ do exercício Exe 3.60.

(a) Determine as matrizes P, L_1', L_2', $\bar{L}$ e L.

(b) Verifique que $PA = LU$.

Obs 3.65

Há uma maneira prática para determinar P, L e U da fatorização PLU de Doolittle? Sim — por exemplo, o "Algoritmo Fatorização PLU de Doolittle" (AFaPLUD), que se apresenta a seguir.

Alg 3.66 — "Algoritmo Fatorização PLU de Doolittle" (AFaPLUD)

Input: matriz $A = [a_{ij}] \in \mathcal{M}_{n\times n}(\mathbb{R})$

Output: matrizes $P = [p_{ij}]$, $L = [\ell_{ij}]$, $U = [u_{ij}] \in \mathcal{M}_{n\times n}(\mathbb{R})$

1 $P \leftarrow I_n$, $L \leftarrow I_n$, $U \leftarrow A$;

2 **for** $k \leftarrow 1$ **to** $n-1$ **do**

3 $\quad i \leftarrow \underset{\bar{i}\in\{k,\ldots,n\}}{\arg\max} |u_{\bar{i}k}|$;

4 $\quad$ **if** $k \neq i$ **then**

5 $\quad\quad p_{k,:} \leftrightarrow p_{i,:}, \quad \ell_{k,1:k-1} \leftrightarrow \ell_{i,1:k-1}, \quad u_{k,k:n} \leftrightarrow u_{i,k:n}$;

6 $\quad$ **for** $j \leftarrow k+1$ **to** n **do**

7 $\quad\quad \ell_{jk} \leftarrow \frac{u_{jk}}{u_{kk}}, \quad u_{j,k:n} \leftarrow u_{j,k:n} - \ell_{jk} u_{k,k:n}$;

Exe 3.67

Seja a matriz invertível $A = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix}$.

(a) Aplique o AFaPLUD à matriz A.

(b) Verifique que $PA = LU$.

Res

(a) **Passo 1**

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix}.$$

Passo 2 $k = 1$

Passo 2.1 $i = 2$: $p_{1,:} \leftrightarrow p_{2,:}$, $\ell_{1,1:0} \leftrightarrow \ell_{2,1:0}$, $u_{1,1:3} \leftrightarrow u_{2,1:3}$

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 2 & 4 \\ 2 & 4 & 1 \end{bmatrix}.$$

Res (cont.)

Passo 2.2 $j = 2$: $\ell_{21} \leftarrow \frac{u_{21}}{u_{11}}$, $u_{2,1:3} \leftarrow u_{2,1:3} - \ell_{21} u_{1,1:3}$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.25 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 1.75 & 3.75 \\ 2 & 4 & 1 \end{bmatrix}.$$

Passo 2.2 $j = 3$: $\ell_{31} \leftarrow \frac{u_{31}}{u_{11}}$, $u_{3,1:3} \leftarrow u_{3,1:3} - \ell_{31} u_{1,1:3}$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.25 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 1.75 & 3.75 \\ 0 & 3.5 & 0.5 \end{bmatrix}.$$

Res (cont.)

Passo 2 $k = 2$

Passo 2.1 $i = 3$: $p_{2,:} \leftrightarrow p_{3,:}$, $\ell_{2,1:1} \leftrightarrow \ell_{3,1:1}$, $u_{2,2:3} \leftrightarrow u_{3,2:3}$

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 3.5 & 0.5 \\ 0 & 1.75 & 3.75 \end{bmatrix}.$$

Passo 2.2 $j = 3$: $\ell_{32} \leftarrow \frac{u_{32}}{u_{22}}$, $u_{3,2:3} \leftarrow u_{3,2:3} - \ell_{32} u_{2,2:3}$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{bmatrix}, U = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 3.5 & 0.5 \\ 0 & 0 & 3.5 \end{bmatrix}.$$

Res (cont.)

(b)

$$PA = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 1 & 2 & 4 \end{bmatrix},$$

$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 1 & 1 \\ 0 & 3.5 & 0.5 \\ 0 & 0 & 3.5 \end{bmatrix} = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 1 & 2 & 4 \end{bmatrix}.$$

Obs 3.68

Está-se agora em condições de resolver sistemas de Cramer $Ax = b$ através de uma fatorização PLU (em particular da fatorização PLU de Doolittle) a tendendo a que $Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb$:

Passo 1 — Resolver o sistema triangular inferior $Ly = Pb$ para y através do algoritmo MeSFeT.

Passo 2 — Resolver o sistema triangular superior $Ux = y$ para x através do algoritmo MeSTaF.

Exe 3.69

Sejam a matriz invertível $A = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 1 \\ 2 & 4 & 1 \end{bmatrix}$ do exercício Exe 3.60 e as matrizes coluna $b_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$ e $b_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. Resolva os sistemas de Cramer $Ax = b_1$ e $Ax = b_2$ recorrendo à fatorização PLU de Doolittle.

Exe 3.70

Sejam a matriz invertível $A = \begin{bmatrix} 1 & 4 & 2 & -2 \\ 5 & 1 & 1 & 0 \\ 2 & 4 & 1 & -3 \\ 1 & -1 & 2 & -2 \end{bmatrix}$ e as matrizes coluna $b_1 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}$ e $b_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}$.

(a) Aplique o AFaPLUD à matriz A.

(b) Verifique que $PA = LU$.

(c) Resolva os sistemas de Cramer $Ax = b_1$ e $Ax = b_2$ recorrendo à fatorização PLU de Doolittle.

Exe 3.71

Sejam a matriz invertível $A = [a_{ij}] \in \mathcal{M}_{4\times 4}(\mathbb{R})$, $a_{ij} = i + j$, e as matrizes coluna $b_1 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}$ e $b_2 = \begin{bmatrix} 1 \\ -2 \\ 3 \\ 1 \end{bmatrix}$.

(a) Determine as matrizes L e U da fatorização LU de Doolittle.

(b) Verifique que $A = LU$.

(c) Resolva os sistemas de Cramer $Ax = b_1$ e $Ax = b_2$ recorrendo à fatorização LU de Doolittle.

(d) Aplique o AFaPLUD à matriz A.

(e) Verifique que $PA = LU$.

(f) Resolva os sistemas de Cramer $Ax = b_1$ e $Ax = b_2$ recorrendo à fatorização PLU de Doolittle.

Exe 3.72

(a) Implemente em MATLAB/Octave o AFaPLUD e compare com os resultados obtidos pela função do MATLAB/Octave `[P,L,U]=lu(A)`.

(b) Implemente em MATLAB/Octave uma função para resolver sistemas de Cramer através da fatorização PLU de Doolittle e compare com os resultados obtidos pela função do MATLAB/Octave `x=A\b`.

Obs 3.73

Seja (S) um sistema de Cramer $Ax = b$. Em muitas situações, os elementos da matriz dos coeficientes A ou do vetor dos termos independentes b estão sujeitos a erros. Estes erros podem resultar de tais elementos serem obtidos a partir de medições (sempre sujeitas a erros) ou de cálculos que originem erros de arredondamento. Assim, pretende-se agora estudar a sensibilidade da solução do sistema (S) face a perturbações quer na matriz A, quer no vetor b.

Def 3.74

⟦sistema bem condicionado, sistema mal condicionado⟧ Seja (S) um sistema de Cramer. Diz-se que (S) é um sistema bem condicionado se pequenas perturbações nos dados (matriz dos coeficientes e/ou vetor dos termos independentes) provocarem pequenas perturbações nos resultados. Caso contrário, diz-se mal condicionado.

Obs 3.75

Comece-se por recordar/introduzir o conceito de norma quer para vetores, quer para matrizes.

Def 3.76

⟦norma⟧ Sejam V um espaço vetorial e a aplicação

$$\begin{array}{rccc} \|\cdot\|: & V & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \|x\|. \end{array}$$

Diz-se que esta aplicação é uma norma se:

(a) $\forall x, y \in V [\|x + y\| \leqslant \|x\| + \|y\|]$.

(b) $\forall x \in V, \forall \alpha \in \mathbb{R} [\|\alpha x\| = |\alpha| \|x\|]$.

(c) $\forall x \in V - \{0_V\} [\|x\| > 0]$.

(d) $\|0_V\| = 0$.

Obs 3.77

(a) A definição de norma generaliza o conceito de comprimento de um vetor.

(b) Não confundir a notação "$\cdot$" para referir "produto interno" e para referir um elemento genérico do domínio na expressão $\|\cdot\|$.

Obs 3.78

Apresentam-se no teorema seguinte normas do espaço vetorial $\mathbb{R}^n$.

Teo 3.79

Seja $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. As seguintes funções são normas de $\mathbb{R}^n$.

Norma um: $\|x\|_1 \stackrel{\text{def}}{=} |x_1| + \cdots + |x_n|$.

Norma dois ou Euclidiana: $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{x_1^2 + \cdots + x_n^2}$.

Norma infinita: $\|x\|_\infty \stackrel{\text{def}}{=} \max_{i \in \{1,\ldots,n\}} |x_i|$.

Exe 3.80

Calcule a norma um, dois e infinita de $x = (1, -2, 1, 0)$.

Res

- $\|x\|_1 = |1| + |-2| + |1| + |0| = 4$.
- $\|x\|_2 = \sqrt{1^2 + (-2)^2 + 1^2 + 0^2} = \sqrt{6}$.
- $\|x\|_\infty = \max(|1|, |-2|, |1|, |0|) = 2$.

Obs 3.81

Pretende-se agora apresentar normas do espaço vetorial $\mathcal{M}_{n\times n}(\mathbb{R})$. Antes disso, define-se raio espetral de uma matriz quadrada.

Def 3.82

⟦raio espetral de uma matriz⟧ Seja a matriz $C \in \mathcal{M}_{n\times n}(\mathbb{R})$. Chama-se raio espetral da matriz C, que se representa por $\rho(C)$, ao máximo dos módulos dos valores próprios de C.

Exe 3.83

Seja C uma matriz quadrada tal que $\lambda(C) = \{-3, 0, 2\}$. Determine o seu raio espetral.

Res

$\rho(C) = \max(|-3|, |0|, |2|\} = 3$.

Teo 3.84

Seja a matriz $A = [a_{ij}] \in \mathcal{M}_{n\times n}(\mathbb{R})$. As seguintes funções são normas de $\mathcal{M}_{n\times n}(\mathbb{R})$.

Norma um: $\|A\|_1 \stackrel{\text{def}}{=} \max_{j \in \{1,\ldots,n\}} \|c_{j,A}\|_1$.

Norma dois ou Euclidiana: $\|A\|_2 \stackrel{\text{def}}{=} \rho(A^\top A)$.

Norma infinita: $\|A\|_\infty \stackrel{\text{def}}{=} \max_{i \in \{1,\ldots,n\}} \|\ell_{i,A}\|_1$.

Exe 3.85

Calcule as normas um e infinita da matriz $A = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 2 & 3 \\ 1 & 3 & -2 \end{bmatrix}$.

Res

- $\|A\|_1 = \max\left(|1| + |-1| + |1|, |0| + |2| + |3|, |2| + |3| + |-2|\right) = \max(3, 5, 7) = 7$.
- $\|A\|_\infty = \max\left(|1| + |0| + |2|, |-1| + |2| + |3|, |1| + |3| + |-2|\right) = \max(3, 6, 6) = 6$.

Def 3.86

⟦número de condição de uma matriz⟧ Seja A uma matriz quadrada invertível. Chama-se número de condição da matriz A na norma p, que se representa por $\text{cond}_p(A)$, a

$$\text{cond}_p(A) \stackrel{\text{def}}{=} \|A\|_p \, \|A^{-1}\|_p.$$

Teo 3.87

Seja A uma matriz quadrada invertível. Então, $\text{cond}(A)_p \geqslant 1$.

Obs 3.88

Seja $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ uma matriz invertível. Recorde que

$$A^{-1} = \frac{1}{|A|}\operatorname{adj}(A) = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Exe 3.89

Seja $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$. Calcule $\operatorname{cond}_1(A)$ e $\operatorname{cond}_\infty(A)$.

Res

Atendendo a que $A^{-1} = -\frac{1}{2}\left[\begin{smallmatrix} 4 & -2 \\ -3 & 1 \end{smallmatrix}\right]$, tem-se que:

- $\operatorname{cond}_1(A) = \|A\|_1 \, \|A^{-1}\|_1 = \max(4,6) \times \max(3.5, 1.5) = 21.$
- $\operatorname{cond}_\infty(A) = \|A\|_\infty \, \|A^{-1}\|_\infty = \max(3,7) \times \max(3,2) = 21.$

Obs 3.90

Apresenta-se agora o teorema que considera perturbações no vetor dos termos independentes.

Teo 3.91

Sejam $\bar{x}$ a solução do sistema de Cramer não-homogéneo $Ax = b$ e $\tilde{x}$ a solução do sistema de Cramer não-homogéneo (perturbado) $Ax = \tilde{b}$. Então, tem-se que

$$\frac{\|\bar{x} - \tilde{x}\|_p}{\|\bar{x}\|_p} \leqslant \operatorname{cond}_p(A) \frac{\|b - \tilde{b}\|_p}{\|b\|_p}.$$

Obs 3.92

O resultado apresentado afirma que "variações relativas" no termo independente aparecem multiplicadas pelo número de condição de A como "variações relativas" na solução do sistema. Note-se que o majorante apresentado pode ser, por vezes, bastante pessimista.

Exe 3.93

Seja (S) o sistema de Cramer cuja matriz dos coeficientes é $A = \left[\begin{smallmatrix} 1 & 2 & 4 \\ 4 & 3 & 1 \\ 2 & 2 & 3 \end{smallmatrix}\right]$ e cujo vetor dos termos independentes é $b = \left[\begin{smallmatrix} 1 \\ 2 \\ 1 \end{smallmatrix}\right]$. Seja, ainda, $(\tilde{S})$ o sistema de Cramer cuja matriz dos coeficientes é A e cujo vetor dos termos independentes é $\tilde{b} = \left[\begin{smallmatrix} 1.1 \\ 2.2 \\ 0.9 \end{smallmatrix}\right]$. Mostre que

$$\frac{\|\bar{x} - \tilde{x}\|_\infty}{\|\bar{x}\|_\infty} \leqslant \operatorname{cond}_\infty(A) \frac{\|b - \tilde{b}\|_\infty}{\|b\|_\infty}.$$

Obs 3.94

Apresenta-se agora o teorema que considera perturbações na matriz dos coeficientes.

Teo 3.95

Sejam $\bar{x}$ a solução do sistema de Cramer não-homogéneo $Ax = b$ e $\tilde{x}$ a solução do sistema de Cramer não-homogéneo (perturbado) $\tilde{A}x = b$. Então, tem-se que

$$\frac{\|\bar{x} - \tilde{x}\|_p}{\|\bar{x}\|_p} \leqslant \operatorname{cond}_p(A) \frac{\|A - \tilde{A}\|_p}{\|A\|_p}.$$

Obs 3.96

O resultado apresentado afirma que "variações relativas" na matriz dos coeficientes aparecem multiplicadas pelo número de condição de A como "variações relativas" na solução do sistema. Note-se que o majorante apresentado pode ser, por vezes, bastante pessimista.

Exe 3.97

Seja (S) o sistema de Cramer cuja matriz dos coeficientes é $A = \begin{bmatrix} 1 & 5 & 10 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{bmatrix}$ e cujo vetor dos termos independentes é $b = \begin{bmatrix} 16 \\ -5 \\ 1 \end{bmatrix}$. Seja, ainda, $(\tilde{S})$ o sistema de Cramer cuja matriz dos coeficientes é $\tilde{A} = \begin{bmatrix} 1 & 5 & 10 \\ 0 & 1 & -6 \\ 0 & 0 & 1.1 \end{bmatrix}$ e cujo vetor dos termos independentes é b. Mostre que

$$\frac{\|\bar{x} - \tilde{x}\|_\infty}{\|\bar{x}\|_\infty} \leqslant \text{cond}_\infty(A) \frac{\|A - \tilde{A}\|_\infty}{\|A\|_\infty}.$$

Obs 3.98

Seja (S) o sistema de Cramer $Ax = b$. Então:

(a) Se $\text{cond}_p(A)$ pequeno, então (S) é um sistema bem condicionado.

(b) Se $\text{cond}_p(A)$ grande, então (S) pode ser um sistema mal condicionado.

Exe 3.99

Seja (S) o sistema de Cramer

$$\begin{cases} x_1 + 0.25x_2 = 1.25 \\ x_1 + \quad x_2 = 2. \end{cases}$$

(a) Calcule $\text{cond}_\infty(A)$.

(b) Indique, justificando, se (S) é um sistema bem condicionado.

Exe 3.100

Seja (S) o sistema de Cramer

$$\begin{cases} 2x_1 + \quad 3x_2 = 1 \\ 2x_1 + 3.0001x_2 = 0.9999. \end{cases}$$

(a) Resolva-o pelo MGPP.

(b) Calcule $\text{cond}_\infty(A)$ e comente os resultados obtidos.

(c) Considere no elemento a_{22} a perturbação $\delta a_{22} = -0.0002$. Resolva este sistema perturbado e comente os resultados obtidos.

Def 4.1

Sejam os pontos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ tais que $x_0 < x_1 < \cdots < x_n$.

(a) ⟦função interpoladora⟧ Diz-se que a função real de variável real g é uma função interpoladora dos $n+1$ pontos dados se

$$g(x_i) = y_i, i = 0, 1, \ldots, n.$$

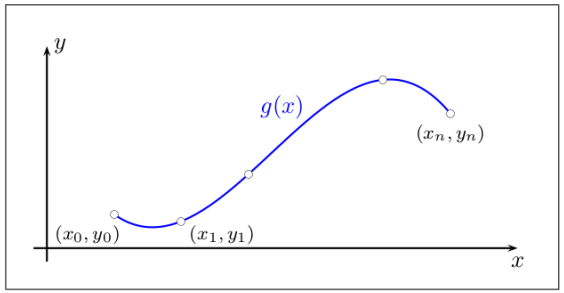
(b) ⟦nós de interpolação⟧ Os valores $x_0, x_1, \ldots, x_n$ designam-se por nós de interpolação.

(c) ⟦valores nodais⟧ Os valores $y_0, y_1, \ldots, y_n$ designam-se por valores nodais.

(d) ⟦tabela matemática⟧ Os $n+1$ pontos dados definem uma tabela matemática.

Obs 4.2

Exemplo de uma função interpoladora g associada à tabela matemática $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$:



Obs 4.3

(a) Importância da interpolação
- Aproximar funções.
- Fundamento de muitos métodos numéricos.

(b) Perante um dado problema de interpolação é necessário escolher a classe de funções interpoladoras a utilizar e a forma de determinar concretamente a função (ou uma função) interpoladora.

(c) Neste curso a classe de funções interpoladoras serão funções polinomiais ($p(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + a_n x^n$) pois:
- As funções polinomiais podem ser calculadas recorrendo a um número finito de operações aritméticas.
- As operações de derivação e primitivação de funções polinomiais são simples.
- As funções polinomiais são funções contínuas e todas as suas derivadas também são funções contínuas.

Teo 4.4

(Existência e unicidade do polinómio interpolador) Seja uma tabela matemática com $n+1$ pontos. Então, existe um e um só polinómio p de grau menor ou igual a n que interpola a tabela matemática.

Obs 4.5

Como calcular o polinómio interpolador da tabela matemática $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$? Uma possibilidade é resolver o sistema de equações lineares

$$\sum_{j=0}^{n} a_j x_i^j = y_i, \quad i = 0, 1, \ldots, n.$$

Esta abordagem não é aconselhável, pois por um lado exige um número elevado de cálculos e por outro o sistema resultante pode ser mal condicionado. Apresenta-se a seguir dois processos para o determinar: a forma de Lagrange e a forma de Newton. Note-se que são dois processos distintos para obter o mesmo polinómio interpolador.

Teo 4.6

(Forma de Lagrange) O polinómio interpolador da tabela matemática $(x_0, y_0), \ldots, (x_n, y_n)$ é

$$p(x) = \sum_{k=0}^{n} L_k(x) y_k.$$

com

$$L_k(x) \stackrel{\text{def}}{=} \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$

Exe 4.7

Seja a tabela matemática $(x_0, y_0) = (0, 1)$, $(x_1, y_1) = (1, 0)$ e $(x_2, y_2) = (2, 3)$.

(a) Determine, considerando a forma de Lagrange, o polinómio interpolador p de grau menor ou igual a 2 que interpola a tabela matemática.

(b) Verifique que p interpola a tabela matemática.

Res

(a) Atendendo a

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 1)(x - 2)}{(0 - 1)(0 - 2)} = \frac{1}{2}(x - 1)(x - 2),$$
$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 0)(x - 2)}{(1 - 0)(1 - 2)} = -x(x - 2),$$
$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 0)(x - 1)}{(2 - 0)(2 - 1)} = \frac{1}{2}x(x - 1),$$

Res (cont.)

tem-se que

$$\begin{aligned} p(x) &= \sum_{k=0}^{2} L_k(x) y_k \\ &= \underbrace{L_0(x) y_0}_{k=0} + \underbrace{L_1(x) y_1}_{k=1} + \underbrace{L_2(x) y_2}_{k=2} \\ &= \frac{1}{2}(x - 1)(x - 2) \times 1 - x(x - 2) \times 0 + \frac{1}{2}x(x - 1) \times 3 \\ &= \frac{1}{2}(x^2 - 3x + 2) + \frac{3}{2}(x^2 - x) \\ &= 2x^2 - 3x + 1. \end{aligned}$$

(b)

$$\begin{aligned} p(0) &= 1 = y_0, \\ p(1) &= 2 - 3 + 1 = 0 = y_1, \\ p(2) &= 8 - 6 + 1 = 3 = y_2. \end{aligned}$$

Obs 4.8

Desvantagens da representação de Lagrange:

(a) é possível obter o polinómio interpolador com menos operações aritméticas.

(b) Passar de k nós de interpolação para $k + 1$: têm que se refazer todos os cálculos.

Teo 4.9

(Forma de Newton) O polinómio interpolador da tabela matemática (x_0, y_0), ..., (x_n, y_n) é

$$\begin{aligned} p(x) = {} & y_0 \\ & + (x - x_0)[x_0, x_1] \\ & + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ & + (x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] \\ & + \cdots \\ & + (x - x_0) \cdots (x - x_{n-1})[x_0, \ldots, x_n], \end{aligned}$$

em que $[x_0, x_1]$, ..., $[x_0, \ldots, x_n]$ obtêm-se a partir da tabela das diferenças divididas

Teo 4.9 (cont.)

$$\begin{array}{c|cccc} x_0 & \boxed{y_0} & \boxed{[x_0, x_1]} & & \\ x_1 & y_1 & [x_1, x_2] & \boxed{[x_0, x_1, x_2]} & \\ x_2 & y_2 & & & \\ \vdots & \vdots & \vdots & \vdots & \cdots \quad \boxed{[x_0, \ldots, x_n]} \\ x_{n-1} & y_{n-1} & [x_{n-1}, x_n] & [x_{n-2}, x_{n-1}, x_n] & \\ x_n & y_n & & & \end{array}$$

Teo 4.9 (cont.)

com

- diferença dividida de primeira ordem

$$[x_i, x_{i+1}] \stackrel{\text{def}}{=} \frac{y_i - y_{i+1}}{x_i - x_{i+1}}.$$

- diferença dividida de segunda ordem

$$[x_i, x_{i+1}, x_{i+2}] \stackrel{\text{def}}{=} \frac{[x_i, x_{i+1}] - [x_{i+1}, x_{i+2}]}{x_i - x_{i+2}}.$$

- ...
- diferença dividida de ordem n

$$[x_i, \ldots, x_{i+n}] \stackrel{\text{def}}{=} \frac{[x_i, \ldots, x_{i+n-1}] - [x_{i+1}, \ldots, x_{i+n}]}{x_i - x_{i+n}}.$$

Exe 4.10

Seja a tabela matemática $(x_0, y_0) = (0, 1)$, $(x_1, y_1) = (1, 0)$ e $(x_2, y_2) = (2, 3)$. Determine, considerando a forma de Newton, o polinómio interpolador p de grau menor ou igual a 2 que interpola a tabela matemática.

Res

Atendendo a

$$\begin{array}{cc|ll} x_0 = 0 & y_0 = \boxed{1} & [x_0, x_1] = \frac{y_0 - y_1}{x_0 - x_1} = \boxed{-1} & [x_0, x_1, x_2] = \frac{[x_0, x_1] - [x_1, x_2]}{x_0 - x_2} = \boxed{2} \\ x_1 = 1 & y_1 = 0 & [x_1, x_2] = \frac{y_1 - y_2}{x_1 - x_2} = 3 & \\ x_2 = 2 & y_2 = 3 & & \end{array}$$

tem-se que

$$\begin{aligned} p(x) &= y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &= 1 + (x - 0) \times (-1) + (x - 0)(x - 1) \times 2 \\ &= 2x^2 - 3x + 1. \end{aligned}$$

Obs 4.11

(a) Note-se que, como a tabela matemática dos exercícios Exe 4.7 e Exe 4.10 é a mesma, o polinómio interpolador, determinado através da forma de Lagrange e da forma de Newton é, obviamente, o mesmo.

(b) No teorema seguinte apresenta-se um majorante do erro do polinómio interpolador num ponto do domínio da função interpolada.

Teo 4.12

Sejam $f \in C^{n+1}([a,b];\mathbb{R})$ e p o polinómio de grau menor ou igual a n que interpola f nos nós de interpolação x_0, x_1, …, x_n pertencentes a $[a,b]$. Seja, ainda, $\bar{x} \in [a,b]$. Então:

$$|f(\bar{x}) - p(\bar{x})| \leqslant \frac{h^{n+1}}{4(n+1)} M,$$

em que

$$M \geqslant \max_{\xi \in [a,b]} |f^{(n+1)}(\xi)|,$$
$$h = \max_{i \in \{0,\ldots,n-1\}} (x_{i+1} - x_i).$$

Exe 4.13

Seja a tabela matemática associada a $f \in C^{\infty}([1,10];\mathbb{R})$

i	0	1	2	3	4	5	6
x_i	1	2	3	4	5	6	10
$y_i = f(x_i)$	0.0000	0.3010	0.4771	0.6021	0.6990	0.7782	1.0000

(a) Calcule uma aproximação a $f(2.75)$ através do polinómio interpolador de grau menor ou igual 2.

(b) Verifique que é válido o majorante apresentado na teorema Teo 4.12 sabendo que $f(x) = \log_{10}(x)$.

Res

(a) Considere-se a forma de Newton. Escolhendo-se os pontos mais próximos de 2.75 tem-se a tabela das diferenças divididas

2	$\boxed{0.3010}$		
		$\boxed{0.1761}$	
3	0.4771		$\boxed{-0.0256}$
		0.1250	
4	0.6021		

vindo

$$\begin{aligned} p(x) &= y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &= 0.3010 + (x-2) \times 0.1761 + (x-2)(x-3) \times (-0.0256), \end{aligned}$$

pelo que

$$f(2.75) \approx p(2.75) = 0.4379.$$

Res

(b) Atendendo a $f(x) = \log_{10}(x)$, tem-se que

$$f^{(1)}(x) = \frac{1}{\ln(10)}\frac{1}{x}, f^{(2)}(x) = -\frac{1}{\ln(10)}\frac{1}{x^2}, f^{(3)}(x) = \frac{1}{\ln(10)}\frac{2}{x^3},$$

pelo que

$$M = \max_{\xi\in[2,4]} |f^{(2+1)}(\xi)| = \max_{\xi\in[2,4]} \frac{1}{\ln(10)}\frac{2}{\xi^3} = \frac{1}{\ln(10)}\frac{2}{2^3} = 0.1086.$$

Como $h = \max(3-2, 4-3) = 1$, tem-se:

$$|f(2.75) - p(2.75)| \leqslant \frac{1^{2+1}}{4(2+1)} \times 0.1086 = 0.0091.$$

Como $f(2.75) = \log_{10}(2.75) = 0.4393$, tem-se que

$$|f(2.75) - p(2.75)| = |0.4393 - 0.4379| = 0.0014 \leqslant 0.0091, \text{ c.q.m.}$$

Exe 4.14

Considere a tabela matemática $(0,1)$, $(1,1)$, $(2,2)$ e $(4,5)$.

(a) Determine, considerando a forma de Lagrange, o polinómio interpolador p de grau menor ou igual a 3 que interpola a tabela matemática.

(b) Determine, considerando a forma de Newton, o polinómio interpolador p de grau menor ou igual a 3 que interpola a tabela matemática.

Exe 4.15

Seja a função real de variável real $f(x) = \int_0^x \exp(t^2)\,\mathrm{d}t$. Suponha que o valor $f(1.15)$ deve ser determinado a partir de uma tabela dos valores $f(1.0)$, $f(1.1)$ e $f(1.2)$. Calcule um majorante para o erro absoluto da aproximação, usando interpolação parabólica ($n = 2$).

Exe 4.16

Seja a tabela matemática associada a $f \in C^\infty([0, \frac{\pi}{2}]; \mathbb{R})$

i	0	1	2	3	4
x_i	0	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{3\pi}{8}$	$\frac{\pi}{2}$
$y_i = f(x_i)$	0	0.3827	0.7071	0.9239	1

(a) Determine, considerando a forma de Lagrange, o polinómio interpolador p de grau menor ou igual a 4 que interpola a tabela matemática.

(b) Determine, considerando a forma de Newton, o polinómio interpolador p de grau menor ou igual a 4 que interpola a tabela matemática.

(c) Calcule uma aproximação a $f(\frac{\pi}{6})$ através do polinómio interpolador de grau menor ou igual a 4.

(d) Relativamente à alínea anterior, verifique que é válido o majorante apresentado na teorema Teo 4.12 sabendo que $f(x) = \mathrm{sen}(x)$.

Exe 4.17

Seja a tabela matemática (x_0, y_0), …, (x_n, y_n).

(a) Implemente em MATLAB/Octave uma função para calcular a tabela das diferenças divididas associada à tabela matemática dada.

(b) Sejam $\bar{x} \in [x_0, x_n]$ e p o polinómio interpolador de grau menor ou igual n da tabela matemática dada. Implemente em MATLAB/Octave uma função para calcular $p(\bar{x})$.

Obs 4.18

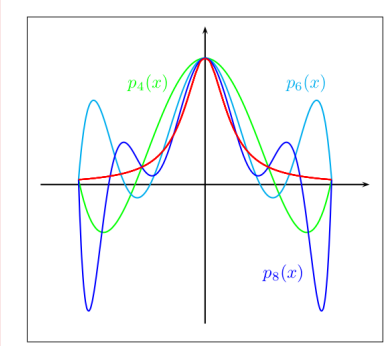
Será que a aproximação é tão melhor quanto maior é o grau do polinómio interpolador? Veja-se o seguinte exemplo.
Sejam a função $f : [-1, 1] \to \mathbb{R}$, definida por $f(x) = \frac{1}{1+25x^2}$ e os seguintes polinómios interpoladores de f:

p_4 com os nós de colocação $x_i = -1 + \frac{i}{4}$, $i = 0, \ldots, 4$.

p_6 com os nós de colocação $x_i = -1 + \frac{i}{3}$, $i = 0, \ldots, 6$.

p_8 com os nós de colocação $x_i = -1 + \frac{i}{4}$, $i = 0, \ldots, 8$.

Obs 4.18 (cont.)



Obs 4.18 (cont.)

Neste exemplo, aumentando o número de nós e mantendo-os equidistantes verifica-se que os polinómios interpoladores apresentam cada vez maiores oscilações. Verifica-se assim que os polinómios interpoladores não se aproximam cada vez mais da função a interpolar como seria desejável.
Uma alternativa é utilizar funções interpoladoras com menos regularidade. Particularmente interessante é a utilização de funções polinomiais por segmentos, isto é, funções que em cada subintervalo sejam definidas por um polinómio, mas que em diferentes subintervalos possam ser definidas por diferentes polinómios. Surge, então, a definição de interpolação polinomial segmentada ou *splines*.

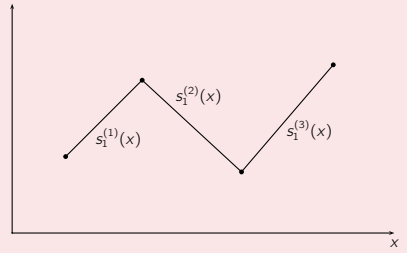
Def 4.19

Uma função $s_\ell : [x_0, x_n] \to \mathbb{R}$ diz-se um *spline* de grau ℓ interpolador da tabela matemática (x_0, y_0), …, (x_n, y_n) se:

- $s_\ell(x_i) = y_i$, $i = 0, \ldots, n$,
- em cada um dos intervalos $[x_{i-1}, x_i]$ $(i = 1, \ldots, n)$ s_ℓ é um polinómio de grau menor ou igual a ℓ (representaremos s_ℓ no sub-intervalo $[x_{i-1}, x_i]$ por $s_\ell^{(i)}$) e
- $s_\ell \in C^{\ell-1}([a, b]; \mathbb{R})$.

Obs 4.20

Spline de grau um ou linear ($\ell = 1$)



$$f(x_i) = s_1(x_i), \qquad i = 0, \ldots, n \qquad \text{(interpola)}$$
$$s_1^{(i)}(x_i) = s_1^{(i+1)}(x_i), \qquad i = 1, \ldots, n-1 \qquad \text{(função contínua)}$$

Teo 4.21

O *spline* linear interpolador da tabela matemática $(x_0, y_0), \ldots, (x_n, y_n)$ é

$$s_1^{(i)}(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}), x \in [x_{i-1}, x_i], i = 1, \ldots, n.$$

Teo 4.22

Sejam $f \in C^2([a, b]; \mathbb{R})$ e s_1 o *spline* linear que interpola f nos nós de colocação $a = x_0 < x_1 < \cdots < x_n = b$. Seja, ainda, $\bar{x} \in [a, b]$. Então,

$$|f(\bar{x}) - s_1(\bar{x})| \leqslant \frac{1}{8} h^2 M,$$

em que

$$M \geqslant \max_{x \in [a,b]} |f^{(2)}(x)|, \quad h = \max_{i \in \{0,\ldots,n-1\}} (x_{i+1} - x_i).$$

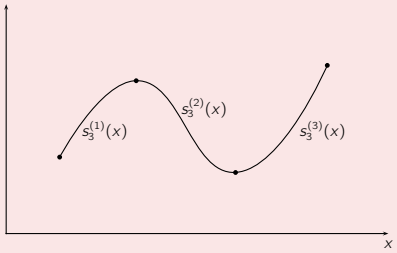
Exe 4.23

Determine o *spline* linear que interpola a tabela matemática

i	0	1	2	3
x_i	-1	0	3	4
y_i	0	1	2	0

Obs 4.24

Spline de grau três ou cúbico ($\ell = 3$)



$$f(x_i) = s_3(x_i), \qquad i = 0, \ldots, n \qquad \text{(interpola)}$$
$$s_3^{(i)}(x_i) = s_3^{(i+1)}(x_i), \qquad i = 1, \ldots, n-1 \qquad \text{(função contínua)}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(i)}(x_i)\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(i+1)}(x_i)\right), \quad i = 1, \ldots, n-1 \quad (1^{\underline{a}} \text{ derivada contínua})$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(i)}(x_i)\right) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(i+1)}(x_i)\right), \quad i = 1, \ldots, n-1 \quad (2^{\underline{a}} \text{ derivada contínua})$$

Teo 4.25

O *spline* cúbico interpolador da tabela matemática (x_0, y_0), …, (x_n, y_n) é

$$\begin{aligned} s_3^{(i)}(x) = {} & \frac{M_{i-1}}{6(x_i - x_{i-1})}(x_i - x)^3 + \frac{M_i}{6(x_i - x_{i-1})}(x - x_{i-1})^3 + \\ & \left(\frac{y_{i-1}}{x_i - x_{i-1}} - \frac{M_{i-1}(x_i - x_{i-1})}{6}\right)(x_i - x) + \\ & + \left(\frac{y_i}{x_i - x_{i-1}} - \frac{M_i(x_i - x_{i-1})}{6}\right)(x - x_{i-1}), \\ & x \in [x_{i-1}, x_i], i = 1, \ldots, n, \end{aligned}$$

Teo 4.25 (cont.)

em que $M_0, \ldots, M_n$ se obtêm a partir da resolução do sistema linear tridiagonal

$$\begin{cases} M_0 = 0 & \text{ponto inicial} \\ \hline (x_i - x_{i-1})M_{i-1} + 2(x_{i+1} - x_{i-1})M_i + (x_{i+1} - x_i)M_{i+1} = \\ \quad \frac{6}{x_{i+1} - x_i}(y_{i+1} - y_i) - \frac{6}{x_i - x_{i-1}}(y_i - y_{i-1}) & \begin{array}{l}\text{pontos interiores} \\ i = 1, \ldots, n-1\end{array} \\ \hline M_n = 0 & \text{ponto final} \end{cases}$$

Exe 4.26

Considere a tabela matemática

i	0	1	2	3	4
x_i	-1	1	2	2.5	3
y_i	-1	2	3	1.5	0

(a) Determine o *spline* cúbico que interpola os pontos da tabela dada.

(b) Verifique que o *spline* calculado na alínea anterior pertence a $C^2([-1, 3]; \mathbb{R})$.

Res

(a) Comece-se por determinar M_1, M_2 e M_3 ($M_0 = 0$ e $M_4 = 0$):

- $i = 1$

$$\begin{aligned} & (x_1 - x_0)M_0 + 2(x_2 - x_0)M_1 + (x_2 - x_1)M_2 = \\ & \quad \frac{6}{x_2 - x_1}(y_2 - y_1) - \frac{6}{x_1 - x_0}(y_1 - y_0) \\ \Leftrightarrow {} & 0 + 2(2 - (-1))M_1 + (2 - 1)M_2 = \\ & \quad \frac{6}{2 - 1}(3 - 2) - \frac{6}{1 - (-1)}(2 - (-1)) \\ \Leftrightarrow {} & 6M_1 + M_2 = -3. \end{aligned}$$

Res (cont.)

- $i = 2$

$$
\begin{aligned}
&(x_2 - x_1)M_1 + 2(x_3 - x_1)M_2 + (x_3 - x_2)M_3 = \\
&\quad \frac{6}{x_3 - x_2}(y_3 - y_2) - \frac{6}{x_2 - x_1}(y_2 - y_1) \\
&\Leftrightarrow (2-1)M_1 + 2(2.5-1)M_2 + (2.5-2)M_3 = \\
&\quad \frac{6}{2.5-2}(1.5-3) - \frac{6}{2-1}(3-2) \\
&\Leftrightarrow M_1 + 3M_2 + 0.5M_3 = -24.
\end{aligned}
$$

Res (cont.)

- $i = 3$

$$
\begin{aligned}
&(x_3 - x_2)M_2 + 2(x_4 - x_2)M_3 + (x_4 - x_3)M_4 = \\
&\quad \frac{6}{x_4 - x_3}(y_4 - y_3) - \frac{6}{x_3 - x_2}(y_3 - y_2) \\
&\Leftrightarrow (2.5-2)M_2 + 2(3-2)M_3 + 0 = \\
&\quad \frac{6}{3-2}(0 - 2.5) - \frac{6}{2.5-2}(1.5-3) \\
&\Leftrightarrow 0.5M_2 + 2M_3 = 0.
\end{aligned}
$$

Tem-se, então, que resolver o sistema

$$
\begin{cases}
6M_1 + M_2 = -3 \\
M_1 + 3M_2 + 0.5M_3 = -24 \\
0.5M_2 + 2M_3 = 0,
\end{cases}
$$

Res (cont.)

ou seja, o sistema de equações lineares cuja matriz dos coeficientes é $A = \begin{bmatrix} 6 & 1 & 0 \\ 1 & 3 & 0.5 \\ 0 & 0.5 & 2 \end{bmatrix}$ e cujo vetor dos termos independentes é $b = \begin{bmatrix} -3 \\ -24 \\ 0 \end{bmatrix}$.
Aplique-se o MGPP para o resolver:

Passo 1 Aplicação do ATEsc-v2:

$$
\left[\begin{array}{ccc|c} 6 & 1 & 0 & -3 \\ 1 & 3 & 0.5 & -24 \\ 0 & 0.5 & 2 & 0 \end{array}\right] \longleftrightarrow \ell_2 \leftarrow \ell_2 - 0.1667\ell_1
$$

$$
\left[\begin{array}{ccc|c} 6 & 1 & 0 & -3 \\ 0 & 2.8333 & 0.5 & -23.5 \\ 0 & 0.5 & 2 & 0 \end{array}\right] \longleftrightarrow \ell_3 \leftarrow \ell_3 - 0.1765\ell_2
$$

$$
\left[\begin{array}{ccc|c} 6 & 1 & 0 & -3 \\ 0 & 2.8333 & 0.5 & -23.5 \\ 0 & 0 & 1.9118 & 4.1471 \end{array}\right].
$$

Res (cont.)

Passo 2 Aplicação do MeSTaF:

- $M_3 = 2.1692$.
- $M_2 = -8.6769$.
- $M_1 = 0.9462$.

O *spline* cúbico é, então, dado por

$$
s_3(x) = \begin{cases} s_3^{(1)}(x), & x \in [-1, 1], \\ s_3^{(2)}(x), & x \in [1, 2], \\ s_3^{(3)}(x), & x \in [2, 2.5], \\ s_3^{(4)}(x), & x \in [2.5, 3], \end{cases}
$$

em que

Res (cont.)

- $i = 1$

$$s_3^{(1)}(x) = \frac{M_0}{6(x_1 - x_0)}(x_1 - x)^3 + \frac{M_1}{6(x_1 - x_0)}(x - x_0)^3 + \left(\frac{y_0}{x_1 - x_0} - \frac{M_0(x_1 - x_0)}{6}\right)(x_1 - x) + \left(\frac{y_1}{x_1 - x_0} - \frac{M_1(x_1 - x_0)}{6}\right)(x - x_0)$$
$$= 0.0788(x + 1)^3 - 0.5(1 - x) + 0.6846(x + 1).$$

Res (cont.)

- $i = 2$

$$s_3^{(2)}(x) = \frac{M_1}{6(x_2 - x_1)}(x_2 - x)^3 + \frac{M_2}{6(x_2 - x_1)}(x - x_1)^3 + \left(\frac{y_1}{x_2 - x_1} - \frac{M_1(x_2 - x_1)}{6}\right)(x_2 - x) + \left(\frac{y_2}{x_2 - x_1} - \frac{M_2(x_2 - x_1)}{6}\right)(x - x_1)$$
$$= 0.1577(2 - x)^3 - 1.4462(x - 1)^3 + 1.8423(2 - x) + 4.4462(x - 1).$$

Res (cont.)

- $i = 3$

$$s_3^{(3)}(x) = \frac{M_2}{6(x_3 - x_2)}(x_3 - x)^3 + \frac{M_3}{6(x_3 - x_2)}(x - x_2)^3 + \left(\frac{y_2}{x_3 - x_2} - \frac{M_2(x_3 - x_2)}{6}\right)(x_3 - x) + \left(\frac{y_3}{x_3 - x_2} - \frac{M_3(x_3 - x_2)}{6}\right)(x - x_2)$$
$$= -2.8923(2.5 - x)^3 + 0.7231(x - 2)^3 + 6.7231(2.5 - x) + 2.8192(x - 2).$$

Res (cont.)

- $i = 4$

$$s_3^{(4)}(x) = \frac{M_3}{6(x_4 - x_3)}(x_4 - x)^3 + \frac{M_4}{6(x_4 - x_3)}(x - x_3)^3 + \left(\frac{y_3}{x_4 - x_3} - \frac{M_3(x_4 - x_3)}{6}\right)(x_4 - x) + \left(\frac{y_4}{x_4 - x_3} - \frac{M_4(x_4 - x_3)}{6}\right)(x - x_3)$$
$$= 0.7231(3 - x)^3 + 2.8192(3 - x).$$

Res (cont.)

(b) • função contínua

$$s_3^{(i)}(x_i) = s_3^{(i+1)}(x_i), i = 1, 2, 3$$

$$s_3^{(1)}(x_1) = 1.9996 \quad s_3^{(2)}(x_1) = 2$$
$$s_3^{(2)}(x_2) = 3 \quad s_3^{(3)}(x_2) = 3$$
$$s_3^{(3)}(x_3) = 1.5 \quad s_3^{(4)}(x_3) = 1.5$$

Res (cont.)

• primeira derivada contínua

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(i)}(x_i)\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(i+1)}(x_i)\right), i = 1, 2, 3$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(1)}(x_1)\right) = 2.1302 \quad \frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(2)}(x_1)\right) = 2.1308$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(2)}(x_2)\right) = -1.7347 \quad \frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(3)}(x_2)\right) = -1.7347$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(3)}(x_3)\right) = -3.3616 \quad \frac{\mathrm{d}}{\mathrm{d}x}\left(s_3^{(4)}(x_3)\right) = -3.3615$$

Res (cont.)

• segunda derivada contínua

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(i)}(x_i)\right) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(i+1)}(x_i)\right), i = 1, 2, 3$$

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(1)}(x_1)\right) = 0.9456 \quad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(2)}(x_1)\right) = 0.9462$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(2)}(x_2)\right) = -8.6772 \quad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(3)}(x_2)\right) = -8.6769$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(3)}(x_3)\right) = 2.1693 \quad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(s_3^{(4)}(x_3)\right) = 2.1693$$

Exe 4.27

Determine o *spline* cúbico que interpola a tabela matemática

i	0	1	2	3
x_i	-1	0	3	4
y_i	0	1	2	0

Exe 4.28

Considere a função real de variável real $f(x) = \exp(x)$.

(a) Aproxime a função f no intervalo $[0, 1]$ através de um *spline* linear baseando-se nos nós de colocação 0, 0.25, 0.75 e 1.

(b) Estime $f(0.5)$ através do *spline* linear da alínea anterior e verifique que é válido o majorante apresentado na teorema Teo 4.22.

(c) Aproxime a função f no intervalo $[0, 1]$ através de um *spline* cúbico baseando-se nos nós de colocação 0, 0.25, 0.75 e 1.

(d) Estime $f(0.5)$ através do *spline* cúbico da alínea anterior.

Exe 4.29

Sejam s_1 o *spline* linear que interpola a tabela matemática (x_0, y_0), ..., (x_n, y_n) e $\bar{x} \in [x_0, x_n]$. Escreva uma função em MATLAB/Octave que calcula $s_1(\bar{x})$.

Exe 4.30

Sejam s_3 o *spline* cúbico que interpola a tabela matemática (x_0, y_0), ..., (x_n, y_n) e $\bar{x} \in [x_0, x_n]$. Escreva uma função em MATLAB/Octave que calcula $s_3(\bar{x})$.

Obs 5.1

(a) Em diversas aplicações é necessário calcular o integral de uma função f para a qual não se conhece uma expressão explícita de uma primitiva, tal primitiva é de obtenção dispendiosa ou quando não se conhece uma expressão para a própria função f. Nestas situações, pode ser utilizada a designada integração numérica que consiste em aproximar o integral

$$I = \int_a^b f(x)\,\mathrm{d}x$$

utilizando apenas valores da função f num conjunto finito de pontos do intervalo $[a, b]$.

Obs 5.1 (cont.)

(b) Métodos a estudar:

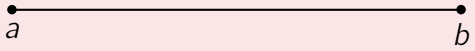
(i) Fórmulas de Newton-Cotes simples — substitui-se f por um polinómio interpolador de f nos pontos igualmente espaçados $x_0 = a$, $x_1 = x_0 + h$, ..., $x_n = b$, com $h = \frac{b-a}{n}$

- $n = 1$: fórmula simples do trapézio.
- $n = 2$: fórmula simples de Simpson.
- $n = 3$: fórmula simples dos três oitavos.

(ii) Fórmulas de Newton-Cotes compostas — aplicação das fórmulas de Newton-Cotes simples várias vezes.

Obs 5.2

- Fórmula simples do trapézio para $\int_a^b f(x)\,\mathrm{d}x$
  - (i) 1 subintervalo.
  - (ii) Posição dos pontos na fórmula simples do trapézio:



  - (iii) Aproximar f por um polinómio de grau menor ou igual a 1 que passa nos pontos $(a, f(a))$ e $(b, f(b))$.

Teo 5.3

Fórmula simples do trapézio: Seja $f \in C^2([a,b];\mathbb{R})$. Então:

$$\exists \xi \in [a,b] : \int_a^b f(x)\,\mathrm{d}x = I_{\mathrm{ts}} + ET_{\mathrm{ts}}$$

em que

$$I_{\mathrm{ts}} = \frac{(b-a)}{2}\left(f(a)+f(b)\right),$$
$$ET_{\mathrm{ts}} = -\frac{(b-a)^3}{12} f^{(2)}(\xi)$$

(I_{ts} é o valor estimado através da fórmula simples do trapézio do integral que se pretende calcular e ET_{ts} é o erro de truncatura desta fórmula).

Exe 5.4

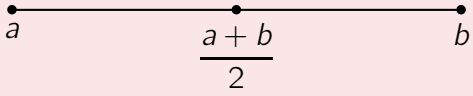
Seja a função real de variável real $f(x) = \exp(x)$. Estime $I = \int_{-2}^{1} f(x)\,\mathrm{d}x$ através da fórmula simples do trapézio (considere 4 casas decimais).

Res

$$I_{\mathrm{ts}} = \int_{-2}^{1} f(x)\,\mathrm{d}x \approx \frac{1-(-2)}{2}\left(f(-2)+f(1)\right)$$
$$= 1.5(0.1353 + 2.7183) = 4.2804.$$

Obs 5.5

- Fórmula simples de Simpson para $\int_a^b f(x)\,\mathrm{d}x$
  - (i) 2 subintervalos igualmente espaçados.
  - (ii) Posição dos pontos na fórmula simples de Simpson:



  - (iii) Aproximar f por um polinómio de grau menor ou igual a 2 que passa nos pontos $(a, f(a))$, $(\frac{a+b}{2}, f(\frac{a+b}{2}))$ e $(b, f(b))$.

Teo 5.6

Fórmula simples de Simpson: Seja $f \in C^4([a,b];\mathbb{R})$. Então:

$$\exists \xi \in [a,b] : \int_a^b f(x)\,\mathrm{d}x = I_{\mathrm{Ss}} + ET_{\mathrm{Ss}}$$

em que

$$I_{\mathrm{Ss}} = \frac{(b-a)}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right),$$
$$ET_{\mathrm{Ss}} = -\frac{(b-a)^5}{2880} f^{(4)}(\xi)$$

(I_{Ss} é o valor estimado através da fórmula simples de Simpson do integral que se pretende calcular e ET_{Ss} é o erro de truncatura desta fórmula).

Exe 5.7

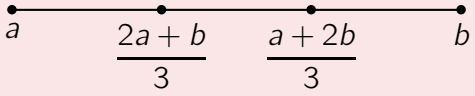
Seja a função real de variável real $f(x) = \exp(x)$. Estime $I = \int_{-2}^{1} f(x)\,\mathrm{d}x$ através da fórmula simples de Simpson (considere 4 casas decimais).

Res

$$I = \int_{-2}^{1} f(x)\,\mathrm{d}x \approx \frac{1-(-2)}{6}\left(f(-2) + 4f\left(\frac{-2+1}{2}\right) + f(1)\right)$$
$$= 0.5(0.1353 + 4 \times 0.6065 + 2.7183) = 2.6398.$$

Obs 5.8

- Fórmula simples dos três oitavos para $\int_a^b f(x)\,\mathrm{d}x$
  - (i) 3 subintervalos igualmente espaçados.
  - (ii) Posição dos pontos na fórmula simples dos três oitavos:



  - (iii) Aproximar f por um polinómio de grau menor ou igual a 3 que passa nos pontos $(a, f(a))$, $\frac{2a+b}{3}, f(\frac{2a+b}{3}))$, $(\frac{a+2b}{3}, f(\frac{a+2b}{3}))$ e $(b, f(b))$.

Teo 5.9

Fórmula simples dos três oitavos: Seja $f \in C^4([a,b];\mathbb{R})$. Então:

$$\exists \xi \in [a,b] : \int_a^b f(x)\,\mathrm{d}x = I_{3/8\mathrm{s}} + ET_{3/8\mathrm{s}}$$

em que

$$I_{3/8\mathrm{s}} = \frac{(b-a)}{8}\left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b)\right),$$
$$ET_{3/8\mathrm{s}} = -\frac{(b-a)^5}{6480} f^{(4)}(\xi)$$

($I_{3/8\mathrm{s}}$ é o valor estimado através da fórmula simples dos três oitavos do integral que se pretende calcular e $ET_{3/8\mathrm{s}}$ é o erro de truncatura desta fórmula).

Exe 5.10

Seja a função real de variável real $f(x) = \exp(x)$. Estime $I = \int_{-2}^{1} f(x)\,\mathrm{d}x$ através da fórmula simples dos três oitavos (considere 4 casas decimais).

Res

$$I = \int_{-2}^{1} f(x)\,\mathrm{d}x \approx \frac{1-(-2)}{8}\left(f(-2) + 3f\left(\frac{2\times(-2)+1}{3}\right) + 3f\left(\frac{-2+2\times 1}{3}\right) + f(1)\right)$$
$$= 0.375(0.1353 + 3\times 0.3679 + 3\times 1 + 2.7183) = 2.6090.$$

Exe 5.11

(a) Relativamente ao exercício Exe 5.4, verifique que é válido o majorante do erro de truncatura obtido a partir do teorema Teo 5.3.

(b) Relativamente ao exercício Exe 5.7, verifique que é válido o majorante do erro de truncatura obtido a partir do teorema Teo 5.6.

(c) Relativamente ao exercício Exe 5.10, verifique que é válido o majorante do erro de truncatura obtido a partir do teorema Teo 5.9.

Obs 5.12

Para diminuir o erro de integração sem aumentar o grau dos polinómios interpoladores utilizam-se regras de integração compostas. Estas consistem em dividir o intervalo $[a, b]$ em subintervalos $[x_0, x_1]$, $[x_1, x_2]$, …, $[x_{n-1}, x_n]$ com $x_0 = a$ e $x_n = b$. Em cada subintervalo $[x_{i-1}, x_i]$, f é interpolada por um polinómio p_i, sendo o integral de f em $[a, b]$ aproximado pela soma dos integrais dos polinómios interpoladores, cada um no subintervalo respetivo, ou seja,

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} p_i(x)\,\mathrm{d}x.$$

Apresentam-se de seguida as fórmulas compostas do trapézio, Simpson e três oitavos.

Teo 5.13

Fórmula composta do trapézio — qualquer número de subintervalos igualmente espaçados: Sejam $f \in C^2([a, b]; \mathbb{R})$ e a tabela matemática $(a = x_0, y_0)$, …, $(b = x_n, y_n)$, tal que $x_{i+1} - x_i = h$, $i = 0, \ldots, n-1$, e $y_i = f(x_i)$, $i = 0, \ldots, n$. Então:

$$\exists \xi \in [a, b] : \int_a^b f(x)\,\mathrm{d}x = I_{\mathrm{tc}} + ET_{\mathrm{tc}}$$

em que

$$I_{\mathrm{tc}} = h\left(\frac{1}{2}y_0 + y_1 + \cdots + y_{n-1} + \frac{1}{2}y_n\right),$$
$$ET_{\mathrm{tc}} = -\frac{(b-a)h^2}{12}f^{(2)}(\xi)$$

(I_{tc} é o valor estimado através da fórmula composta do trapézio do integral que se pretende calcular e ET_{tc} é o erro de truncatura desta fórmula).

Exe 5.14

Seja a função real de variável real $f(x) = \exp(-x^2)$. Estime $\int_0^1 f(x)\,\mathrm{d}x$ através da fórmula composta do trapézio com 6 subintervalos (considere 4 casas decimais).

Res

Atendendo a $h = \frac{1-0}{6} = 0.1667$, tem-se a tabela matemática

i	0	1	2	3	4	5	6
x_i	0	0.1667	0.3333	0.5	0.6677	0.8333	1
$y_i = f(x_i)$	1	0.9726	0.8948	0.7788	0.6412	0.4994	0.3679

Tem-se, então:

$$\int_0^1 f(x)\,\mathrm{d}x \approx h\left(\frac{1}{2}y_0 + y_1 + y_2 + y_3 + y_4 + y_5 + \frac{1}{2}y_6\right) = 0.7451.$$

Exe 5.15

Determine o maior passo h da regra do trapézio composta para aproximar $I = \int_0^2 \frac{1}{1+x}\,\mathrm{d}x$ por forma a que o erro de truncatura não exceda em valor absoluto 5×10^{-5}.

Res

Atendendo a

$$f(x) = \frac{1}{1+x}, f^{(1)}(x) = -\frac{1}{(1+x)^2}, f^{(2)}(x) = \frac{2}{(1+x)^3},$$

tem-se que

$$|ET_{\mathrm{tc}}| \leqslant \frac{(b-a)h^2}{12} \max_{\xi\in[0,2]} |f^{(2)}(\xi)| = \frac{(2-0)h^2}{12}\frac{2}{(1+0)^3} = 0.3333h^2,$$

vindo

Res (cont.)

$$\begin{aligned} &0.3333h^2 \leqslant 5 \times 10^{-5} \Leftrightarrow h^2 \leqslant 1.5002 \times 10^{-4} \Leftrightarrow h \leqslant 0.0122 \\ &\Leftrightarrow \frac{2-0}{n} \leqslant 0.0122 \Leftrightarrow n \geqslant \frac{2}{0.0122} \Leftrightarrow n \geqslant 163.93 \\ &\Rightarrow n = 164. \end{aligned}$$

Assim, o maior passo h para a aplicação da regra do trapézio composta por forma a que o erro de truncatura não exceda em valor absoluto 5×10^{-5} é, com 4 casas decimais,

$$h = \frac{2-0}{164} = 0.0122.$$

Teo 5.16

Fórmula composta de Simpson — número par de subintervalos igualmente espaçados: Sejam $f \in C^4([a,b];\mathbb{R})$ e a tabela matemática $(a = x_0, y_0), \ldots, (x_n, b = y_n)$, tal que $x_{i+1} - x_i = h$, $i = 0, \ldots, n-1$, e $y_i = f(x_i)$, $i = 0, \ldots, n$. Então:

$$\exists \xi \in [a,b] : \int_a^b f(x)\,\mathrm{d}x = I_{\mathrm{Sc}} + ET_{\mathrm{Sc}}$$

em que

$$\begin{aligned} I_{\mathrm{Sc}} &= \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n\right), \\ ET_{\mathrm{Sc}} &= -\frac{(b-a)h^4}{180} f^{(4)}(\xi) \end{aligned}$$

(I_{Sc} é o valor estimado através da fórmula composta de Simpson do integral que se pretende calcular e ET_{Sc} é o erro de truncatura desta fórmula).

Exe 5.17

Seja a função real de variável real $f(x) = \exp(-x^2)$. Estime $\int_0^1 f(x)\,\mathrm{d}x$ através da fórmula compostas de Simpson com 6 subintervalos (considere 4 casas decimais).

Res

Atendendo a $h = \frac{1-0}{6} = 0.1667$, tem-se a tabela matemática

i	0	1	2	3	4	5	6
x_i	0	0.1667	0.3333	0.5	0.6677	0.8333	1
$y_i = f(x_i)$	1	0.9726	0.8948	0.7788	0.6412	0.4994	0.3679

Tem-se, então:

$$\int_0^1 f(x)\,\mathrm{d}x \approx \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + 4y_5 + y_6\right) = 0.7468.$$

Teo 5.18

Fórmula composta dos três oitavos — número múltiplo de três de subintervalos igualmente espaçados: Sejam $f \in C^4([a,b];\mathbb{R})$ e a tabela matemática $(a = x_0, y_0)$, …, $(b = x_n, y_n)$, tal que $x_{i+1} - x_i = h$, $i = 0, \ldots, n-1$, e $y_i = f(x_i)$, $i = 0, \ldots, n$. Então:

$$\exists \xi \in [a,b] : \int_a^b f(x)\,\mathrm{d}x = I_{3/8\mathrm{c}} + ET_{3/8\mathrm{c}}$$

em que

$$I_{3/8\mathrm{c}} = \frac{3h}{8}\left(y_0 + 3y_1 + 3y_2 + 2y_3 + \cdots + 2y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n\right),$$

$$ET_{3/8\mathrm{c}} = -\frac{(b-a)h^4}{80} f^{(4)}(\xi)$$

($I_{3/8\mathrm{c}}$ é o valor estimado através da fórmula composta dos três oitavos do integral que se pretende calcular e $ET_{3/8\mathrm{c}}$ é o erro de truncatura desta fórmula).

Exe 5.19

Seja a função real de variável real $f(x) = \exp(-x^2)$. Estime $\int_0^1 f(x)\,\mathrm{d}x$ através da fórmula compostas dos três oitavos com 6 subintervalos (considere 4 casas decimais).

Res

Atendendo a $h = \frac{1-0}{6} = 0.1667$, tem-se a tabela matemática

i	0	1	2	3	4	5	6
x_i	0	0.1667	0.3333	0.5	0.6677	0.8333	1
$y_i = f(x_i)$	1	0.9726	0.8948	0.7788	0.6412	0.4994	0.3679

Tem-se, então:

$$\int_0^1 f(x)\,\mathrm{d}x \approx \frac{3h}{8}\left(y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + y_6\right) = 0.7468.$$

Obs 5.20

Extensão a intervalos não constantes: começar por dividir o intervalo $[x_0, x_n]$ em subintervalos com espaçamentos constantes, podendo-se depois aplicar as fórmulas de Newton-Cotes, simples ou compostas, a cada um daqueles subintervalos.

Exe 5.21

Considere a tabela matemática

i	0	1	2	3	4	5	6	7	8	9	10	11	12
x_i	0.0	0.1	0.2	0.3	0.5	0.7	0.9	1.1	1.3	1.4	1.5	1.6	1.7
$y_i = f(x_i)$	1.00	1.11	1.23	1.38	1.61	1.82	1.97	2.06	2.25	2.38	2.47	2.61	2.72

Estime $\int_0^{1.7} f(x)\,\mathrm{d}x$.

Exe 5.22

Seja a função real de variável real $f(x) = \dfrac{1}{1+x^2}$.

(a) Estime $I = \int_{-4}^{4} f(x)\,\mathrm{d}x$ através das fórmulas compostas do trapézio e Simpson com 4 e 6 subintervalos (considere seis casas decimais).

(b) Comente os resultados obtidos, dado que o valor exato, com seis casas decimais corretas, de I é 2.651635.

Exe 5.23

Determine o maior passo h da regra de Simpson composta para aproximar $I = \int_0^2 \frac{1}{1+x}\,\mathrm{d}x$ por forma a que o erro de truncatura não exceda em valor absoluto 5×10^{-5}.

Exe 5.24

Determine uma aproximação ao valor do integral definido

$$I = \int_0^1 \left(x^2 + \frac{1}{x+1}\right)\mathrm{d}x,$$

através da fórmula de Simpson com um erro de truncatura, em valor absoluto, inferior a 5×10^{-4}.

Exe 5.25

Seja a função real de variável real

$$f(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t.$$

(a) Estime $f(0.5)$ através

(i) da fórmula do trapézio considerando os pontos 0.0, 0.1, 0.2, 0.3, 0.4 e 0.5.

(ii) da fórmula de Simpson + três oitavos considerando os pontos 0.0, 0.1, 0.2, 0.3, 0.4 e 0.5.

(b) Sabendo que $f(0.5) = 0.5205$, comente os resultados obtidos na alínea anterior.

(c) Determine um majorante do erro de truncatura cometido na alínea (a.i).

Exe 5.26

(a) Sejam f uma função real de variável real e a e b números reais tais que $a < b$. Escreva uma função em MATLAB que estima $\int_a^b f(x)\,\mathrm{d}x$ através da fórmula composta do trapézio com n subintervalos sem recorrer aos comandos `while`, `for` e `repeat` e sem o recurso a funções recursivas.

(b) Escreva uma *script* em MATLAB para resolver o exercício Exe 5.14 recorrendo à função desenvolvida na alínea anterior.

Exe 5.27

(a) Sejam f uma função real de variável real e a e b números reais tais que $a < b$. Escreva uma função em MATLAB que estima $\int_a^b f(x)\,\mathrm{d}x$ através da fórmula composta de Simpson com n subintervalos (n tem que ser um número par) sem recorrer aos comandos `while`, `for` e `repeat` e sem o recurso a funções recursivas.

(b) Escreva uma *script* em MATLAB para resolver o exercício Exe 5.17 recorrendo à função desenvolvida na alínea anterior.

Exe 5.28

(a) Sejam f uma função real de variável real e a e b números reais tais que $a < b$. Escreva uma função em MATLAB que estima $\int_a^b f(x)\,\mathrm{d}x$ através da fórmula composta dos três oitavos com n subintervalos (n tem que ser um múltiplo de 3) sem recorrer aos comandos `while`, `for` e `repeat` e sem o recurso a funções recursivas.

(b) Escreva uma *script* em MATLAB para resolver o exercício Exe 5.19 recorrendo à função desenvolvida na alínea anterior.