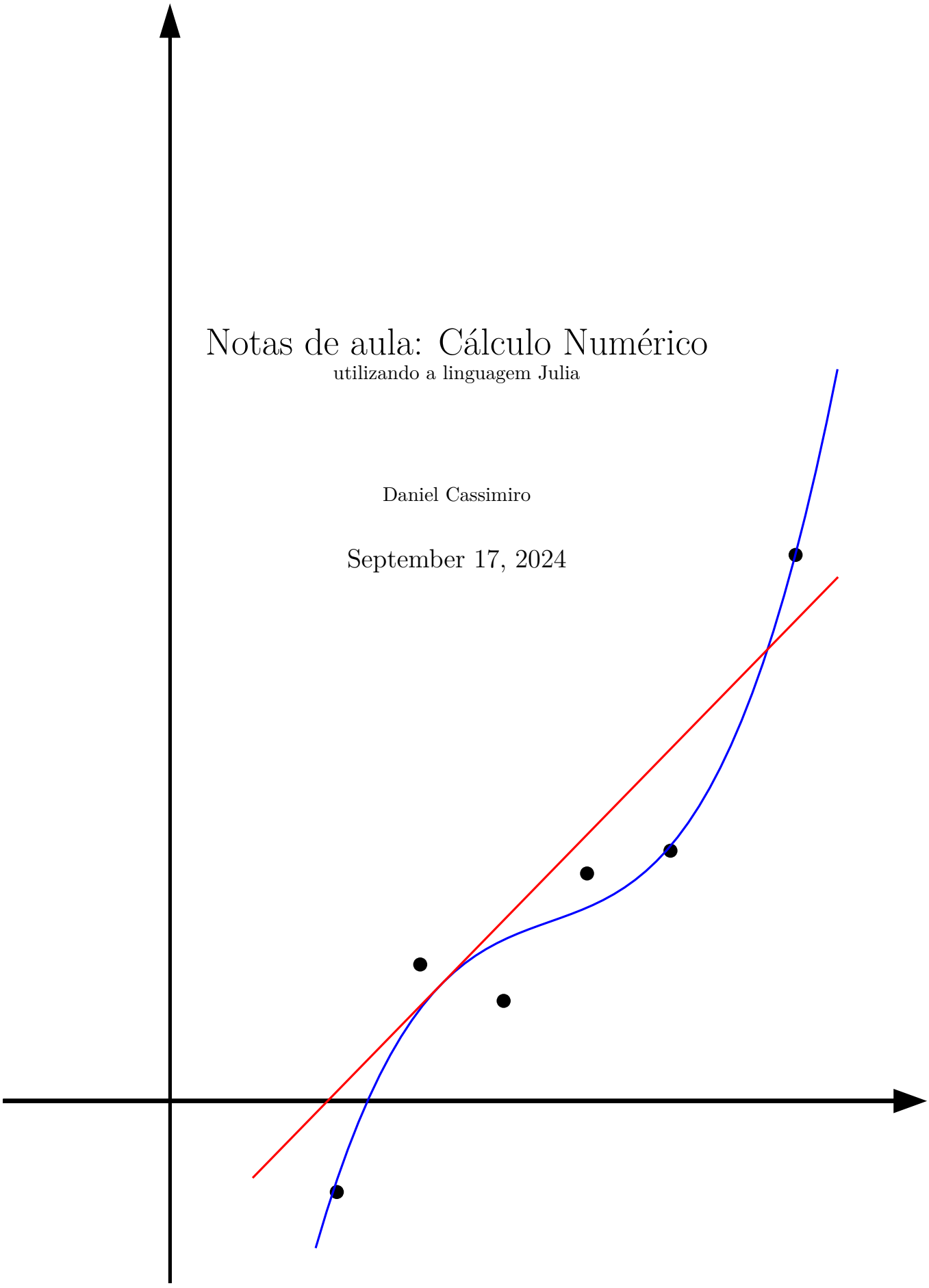
Notas de aula: Cálculo Numérico

utilizando a linguagem Julia

Daniel Cassimiro

September 17, 2024



Licença

Contents

Chapter 1

Introdução

Cálculo numérico é a disciplina que estuda as técnicas para a solução aproximada de problemas matemáticos. Estas técnicas são de natureza analítica e computacional. As principais preocupações normalmente envolvem exatidão e desempenho.

Aliado ao aumento contínuo da capacidade de computação disponível, o desenvolvimento de métodos numéricos tornou a simulação computacional de problemas matemáticos uma prática usual nas mais diversas áreas científicas e tecnológicas. As então chamadas simulações numéricas são constituídas de um arranjo de vários esquemas numéricos dedicados a resolver problemas específicos como, por exemplo: resolver equações algébricas, resolver sistemas de equações lineares, interpolar e ajustar pontos, calcular derivadas e integrais, resolver equações diferenciais ordinárias etc. Neste livro, abordamos o desenvolvimento, a implementação, a utilização e os aspectos teóricos de métodos numéricos para a resolução desses problemas.

Trabalharemos com problemas que abordam aspectos teóricos e de utilização dos métodos estudados, bem como com problemas de interesse na engenharia, na física e na matemática aplicada.

A necessidade de aplicar aproximações numéricas decorre do fato de que esses problemas podem se mostrar intratáveis se dispomos apenas de meios puramente analíticos, como aqueles estudados nos cursos de cálculo e álgebra linear. Por exemplo, o teorema de Abel-Ruffini nos garante que não existe uma fórmula algébrica, isto é, envolvendo apenas operações aritméticas e radicais, para calcular as raízes de uma equação polinomial de qualquer grau, mas apenas casos particulares:

- Simplesmente isolar a incógnita para encontrar a raiz de uma equação do primeiro grau;
- Fórmula de Bhaskara para encontrar raízes de uma equação do segundo grau;
- Fórmula de Cardano para encontrar raízes de uma equação do terceiro grau;
- Existe expressão para equações de quarto grau;

- Casos simplificados de equações de grau maior que 4 onde alguns coeficientes são nulos também podem ser resolvidos.

Equações não polinomiais podem ser ainda mais complicadas de resolver exatamente, por exemplo:

$$\cos(x) = x \qquad \text{ou} \qquad xe^x = 10 \tag{1.1}$$

A maioria dos problemas envolvendo fenômenos reais produzem modelos matemáticos cuja solução analítica é difícil (ou impossível) de obter, mesmo quando provamos que a solução existe. Nesse curso propomos calcular aproximações numéricas para esses problemas, que apesar de, em geral, serem diferentes da solução exata, mostraremos que elas podem ser bem próximas.

Para entender a construção de aproximações é necessário estudar como funciona a aritmética implementada nos computadores e erros de arredondamento. Como computadores, em geral, usam uma base binária para representar números, começaremos falando em mudança de base.

Chapter 2

Representação de números e aritmética de máquina

Neste capítulo, abordaremos formas de representar números reais em computadores. Iniciamos com uma discussão sobre representação posicional e mudança de base. Então, enfatizaremos a representação de números com quantidade finita de dígitos, mais especificamente, as representações de números inteiros, ponto fixo e ponto flutuante em computadores.

A representação de números e a aritmética em computadores levam aos chamados erros de arredondamento e de truncamento. Ao final deste capítulo, abordaremos os efeitos do erro de arredondamento na computação científica.

2.1 Sistema de numeração e mudança de base

Usualmente, utilizamos o sistema de numeração decimal, isto é, base 10, para representar números. Esse é um sistema de numeração em que a posição do algarismo indica a potência de 10 pela qual seu valor é multiplicado.

Exemplo 2.1.1. O número 293 é decomposto como

$$\begin{aligned} 293 &= 2 \text{ centenas} + 9 \text{ dezenas } + 3 \text{ unidades} \\ &= 2 \cdot 10^2 + 9 \cdot 10^1 + 3 \cdot 10^0. \end{aligned} \tag{2.1}$$

O sistema de numeração posicional também pode ser usado com outras bases. Vejamos a seguinte definição.

Definição 2.1.1 (Sistema de numeração de base b)**.** *Dado um número natural $b > 1$ e o conjunto de símbolos $\{\pm, \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{b-1}\}$[1], a sequência de símbolos*

$$(d_n d_{n-1} \cdots d_1 d_0, d_{-1} d_{-2} \cdots)_b \tag{2.2}$$

[1]Para $b > 10$, veja a Observação 2.1.1.

representa o número positivo

$$d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots \tag{2.3}$$

Para representar números negativos usamos o símbolo $-$ à esquerda do numeral[2].

Observação 2.1.1 ($b \geq 10$)**.** Para sistemas de numeração com base $b \geq 10$ é usual utilizar as seguintes notações:

- No sistema de numeração decimal ($b = 10$), costumamos representar o número sem os parênteses e o subíndice, ou seja,

$$\pm d_n d_{n-1} \ldots d_1 d_0{,}d_{-1} d_{-2} \ldots := \pm(d_n d_{n-1} \ldots d_1 d_0{,}d_{-1} d_{-2} \ldots)_{10}. \tag{2.4}$$

- Se $b > 10$, usamos as letras $A, B, C, \cdots$ para denotar os algarismos: $A = 10$, $B = 11$, $C = 12$, $D = 13$, $E = 14$, $F = 15$.

Exemplo 2.1.2 (Sistema binário)**.** O sistema de numeração em base dois é chamado de binário e os algarismos binários são conhecidos como *bits* (do inglês **binary digits**). Um *bit* pode assumir dois valores distintos: 0 ou 1. Por exemplo:

$$\begin{aligned} x &= (1001{,}101)_2 \\ &= 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} \\ &= 8 + 0 + 0 + 1 + 0{,}5 + 0 + 0{,}125 = 9{,}625. \end{aligned} \tag{2.5}$$

Ou seja, $(1001{,}101)_2$ é igual a 9,625 no sistema decimal.

Em `Julia` podemos converter o número $(1001{,}101)_2$ para a base decimal computando

```
julia> 1*2^3 + 0*2^2 + 0*2^1 + 1*2^0 + 1*2^-1 + 0*2^-2 + 1*2^-3
9.625
```

Exemplo 2.1.3 (Sistema quaternário)**.** No sistema quaternário a base b é igual a 4 e, portanto, temos o seguinte conjunto de algarismos $\{0, 1, 2, 3\}$. Por exemplo:

$$(301{,}2)_4 = 3 \cdot 4^2 + 0 \cdot 4^1 + 1 \cdot 4^0 + 2 \cdot 4^{-1} = 49{,}5. \tag{2.6}$$

Verifique no computador!

Exemplo 2.1.4 (Sistema octal)**.** No sistema octal a base é $b = 8$. Por exemplo:

$$\begin{aligned} (1357{,}24)_8 &= 1 \cdot 8^3 + 3 \cdot 8^2 + 5 \cdot 8^1 + 7 \cdot 8^0 + 2 \cdot 8^{-1} + 4 \cdot 8^{-2} \\ &= 512 + 192 + 40 + 7 + 0{,}25 + 0{,}0625 = 751{,}3125. \end{aligned} \tag{2.7}$$

Verifique no computador!

[2] O uso do símbolo $+$ é opcional na representação de números positivos.

Exemplo 2.1.5 (Sistema hexadecimal)**.** O sistema de numeração cuja base é $b = 16$ é chamado de sistema hexadecimal. Neste, temos o conjunto de algarismos $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$. Convertendo o número $(E2AC)_{16}$ para a base 10 temos

$$\begin{aligned}(E2AC)_{16} &= 14 \cdot 16^3 + 2 \cdot 16^2 + 10 \cdot 16^1 + 12 \cdot 16^0 \\ &= 57344 + 512 + 160 + 12 = 58028.\end{aligned} \tag{2.8}$$

Verifique no computador!

Observação 2.1.2. A linguagem `Julia` tem prefixos para representar números nas bases 2, 8 e 10. Por exemplo, temos:

```
>>> print(0b1001) # bin -> dec
9
>>> print(0o157) # oct -> dec
111
>>> print(0xbeba) # hex -> dec
48826
>>> string(9,base=2) # dec -> bin
1001
```

Também é possível usar a função int() para interpretar a representação de um inteiro em uma base com b entre 2 e 36. Por exemplo, temos:

```
>>> print(int('1001', 2)) # Base 2
9
>>> print(int('1001', 5)) # Base 5
126
>>> print(int('ABCD', 20)) # Base 20
84653
>>> print(int('zz', 36)) # Base 36
1295
```

Nos exemplos acima vimos como converter números representados em um sistema de numeração de base b para o sistema decimal. Agora, vamos estudar como fazer o processo inverso. Isto é, dado um número decimal $(X)_{10}$ queremos escrevê-lo em uma outra base b, isto é, queremos obter a seguinte representação:

$$\begin{aligned}(X)_{10} &= (d_n d_{n-1} \cdots d_0, d_{-1} \cdots)_b \\ &= d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots\end{aligned} \tag{2.9}$$

Separando as partes inteira e fracionária de X, isto é, $X = X^{\mathrm{i}} + X^{\mathrm{f}}$, temos

$$X^{\mathrm{i}} = d_n \cdot b^n + \cdots + d_{n-1} b^{n-1} \cdots + d_1 \cdot b^1 + d_0 \cdot b^0 \tag{2.10}$$

e

$$X^{\mathrm{f}} = \frac{d_{-1}}{b^1} + \frac{d_{-2}}{b^2} + \cdots \tag{2.11}$$

Nosso objetivo é determinar os algarismos $\{d_n, d_{n-1}, ...\}$.

Primeiramente, vejamos como tratar a parte inteira X^{i}. Calculando o quociente de X^{i} por b, temos:

$$\frac{X^{\mathrm{i}}}{b} = \frac{d_0}{b} + d_1 + d_2 \cdot b^1 + \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}. \tag{2.12}$$

Observe que d_0 é o resto da divisão de X^{i} por b, pois $d_1 + d_2 \cdot b^1 + \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ é inteiro e $\frac{d_0}{b}$ é uma fração com $d_0 < b$. Da mesma forma, o resto da divisão de $d_1 + d_2 \cdot b^1 + \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ por b é d_1. Ou seja, repetindo este processo encontramos os algarismos d_0, d_1, d_2, ..., d_n.

Vamos, agora, converter a parte fracionária X^{f} do número decimal X para o sistema de base b. Multiplicando X^{f} por b, temos

$$bX^{\mathrm{f}} = d_{-1} + \frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots \tag{2.13}$$

Observe que a parte inteira desse produto é d_{-1} e $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ é a parte fracionária. Quando multiplicamos $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ por b novamente, encontramos d_{-2}. Repetindo este processo encontramos os demais algarismos.

Exemplo 2.1.6. Vamos converter o número 9,625 para a base binária ($b = 2$). Primeiramente, decompomos 9,625 na soma de suas partes inteira e fracionária.

$$9{,}625 = 9 + 0{,}625. \tag{2.14}$$

Conversão da parte inteira. Para converter a parte inteira, fazemos sucessivas divisões por $b = 2$ obtendo

$$\begin{aligned} 9 &= 4 \cdot 2 + 1 && (2.15)\\ &= (2 \cdot 2 + 0) \cdot 2 + 1 && (2.16)\\ &= 2^3 + 1. && (2.17) \end{aligned}$$

Ou seja, temos que $9 = (1001)_2$.

Em `Julia`, podemos usar os comandos `int` (truncamento) e a operação `%` (resto da divisão) para computar esta conversão da seguinte forma

```
>>> x = 9
>>> d0 = x%2; x = int(x/2); print(f"d0 = {d0}, x = {x}")
d0 = 1, x = 4
```

```
>>> d1 = x%2; x = int(x/2); print(f"d1 = {d1}, x = {x}")
d1 = 0, x = 2
>>> d2 = x%2; x = int(x/2); print(f"d2 = {d2}, x = {x}")
d2 = 0, x = 1
>>> d3 = x%2; x = int(x/2); print(f"d3 = {d3}, x = {x}")
d3 = 1, x = 0
```

Conversão da parte fracionária. Para converter a parte fracionária, fazemos sucessivas multiplicações por $b = 2$ obtendo

$$0{,}625 = 1{,}25 \cdot 2^{-1} = 1 \cdot 2^{-1} + 0{,}25 \cdot 2^{-1} \tag{2.18}$$
$$= 1 \cdot 2^{-1} + (0{,}5 \cdot 2^{-1}) \cdot 2^{-1} = 1 \cdot 2^{-1} + 0{,}5 \cdot 2^{-2} \tag{2.19}$$
$$= 1 \cdot 2^{-1} + (1 \cdot 2^{-1}) \cdot 2^{-2} = 1 \cdot 2^{-1} + 1 \cdot 2^{-3}. \tag{2.20}$$

Ou seja, temos que $0{,}625 = (0{,}101)_2$.

No `GNU Octave`, podemos computar esta conversão da parte fracionária da seguinte forma

```
>> x = 0.625
x =  0.62500
>> d = fix(2*x), x = 2*x - d
d =  1
x =  0.25000
>> d = fix(2*x), x = 2*x - d
d = 0
x =  0.50000
>> d = fix(2*x), x = 2*x - d
d =  1
x = 0
```

Conclusão. Da conversão das partes inteira e fracionária de 9,625, obtemos $9 = (1001)_2$ e $0{,}625 = (0{,}101)_2$. Logo, concluímos que $9{,}625 = (1001{,}101)_2$.

Observação 2.1.3. O `GNU Octave` oferece algumas funções para a conversão de números inteiros em base decimal para uma base dada. Por exemplo, temos:

```
>> dec2base(9,2)
ans = 1001
>> dec2base(111,8)
ans = 157
>> dec2base(48826,16)
ans = BEBA
```

Observação 2.1.4. Uma maneira de converter um número dado em uma base b_1 para uma base b_2 é fazer em duas partes: primeiro converter o número dado na base b_1 para base decimal e depois converter para a base b_2.

Exercícios resolvidos

ER 2.1.1. Obtenha a representação do número $125{,}58\overline{3}$ na base 6.

Solução. Decompomos $125{,}58\overline{3}$ nas suas partes inteira 125 e fracionária $0{,}58\overline{3}$. Então, convertemos cada parte.

Conversão da parte inteira. Vamos escrever o número 125 na base 6. Para tanto, fazemos sucessivas divisões por 6 como segue:

$$\begin{aligned} 125 &= 20 \cdot 6 + 5 \quad (125 \text{ dividido por } 6 \text{ é igual a } 20 \text{ e resta } 5) \\ &= (3 \cdot 6 + 2) \cdot 6 + 5 = 3 \cdot 6^2 + 2 \cdot 6 + 5, \end{aligned} \tag{2.21}$$

logo $125 = (325)_6$.

Estes cálculos podem ser feitos no `GNU Octave` com o auxílio das funções `mod` e `fix`. A primeira calcula o resto da divisão entre dois números, enquanto que a segunda retorna a parte inteira de um número dado. No nosso exemplo, temos:

```
>> x = 125
x =  125
>> d = mod(x,6), x = fix(x/6)
d =  5
x =  20
>> d = mod(x,6), x = fix(x/6)
d =  2
x =  3
>> d = mod(x,6), x = fix(x/6)
d =  3
x = 0
```

Verifique!

Conversão da parte fracionária. Para converter $0{,}58\overline{3}$ para a base 6, fazemos sucessivas multiplicações por 6 como segue:

$$\begin{aligned} 0{,}58\overline{3} &= 3{,}5 \cdot 6^{-1} \quad (0{,}58\overline{3} \text{ multiplicado por } 6 \text{ é igual a } 3{,}5) \\ &= 3 \cdot 6^{-1} + 0{,}5 \cdot 6^{-1} \\ &= 3 \cdot 6^{-1} + (3 \cdot 6^{-1}) \cdot 6^{-1} \\ &= 3 \cdot 6^{-1} + 3 \cdot 6^{-2}, \end{aligned} \tag{2.22}$$

logo $0{,}58\overline{3} = (0{,}33)_6$.

No `GNU Octave`, podemos computar esta conversão da parte fracionária da seguinte forma

```
>> x = 0.58 + 1/3/100
x =  0.58333
>> d = fix(6*x), x = 6*x - d
d =  3
x =  0.50000
>> x = 0.5 #isso é realmente necessário?
x =  0.50000
>> d = fix(6*x), x = 6*x - d
d =  3
x = 0
```

◊

ER 2.1.2. Obtenha a representação na base 4 do número $(101{,}01)_2$.

Solução. Começamos convertendo $(101{,}01)_2$ para a base decimal:

$$(101{,}01)_2 = 1 \cdot 2^2 + 1 \cdot 2^0 + 1 \cdot 2^{-2} = 5{,}25. \tag{2.23}$$

Então, convertemos 5,25 para a base 4. Para sua parte inteira, temos

$$5 = 1 \cdot 4 + 1 = (11)_4. \tag{2.24}$$

Para sua parte fracionária, temos

$$0{,}25 = 1 \cdot 4^{-1} = (0{,}1)_4. \tag{2.25}$$

Logo, $(101{,}01)_2 = (11{,}1)_4$. Verifique estas contas no computador! ◊

Exercícios

E 2.1.1. Converta para base decimal cada um dos seguintes números:

a) $(100)_2$

b) $(100)_3$

c) $(100)_b$

d) $(12)_5$

e) $(AA)_{16}$

f) $(7{,}1)_8$

g) $(3{,}12)_5$

E 2.1.2. Escreva os números abaixo na base decimal.

a) $(25{,}13)_8$

b) $(101{,}1)_2$

c) $(12F{,}4)_{16}$

d) $(11{,}2)_3$

E 2.1.3. Escreva o número 5,5 em base binária.

E 2.1.4. Escreva o número 17,109375 em base hexadecimal ($b = 16$).

E 2.1.5. Escreva cada número decimal na base b.

a) $7{,}\overline{6}$ na base $b = 5$

b) $29{,}1\overline{6}$ na base $b = 6$

E 2.1.6. Escreva $(12.4)_8$ em base decimal e binária.

E 2.1.7. Escreva cada número dado para a base b.

a) $(45{,}1)_8$ para a base $b = 2$

b) $(21{,}2)_8$ para a base $b = 16$

c) $(1001{,}101)_2$ para a base $b = 8$

d) $(1001{,}101)_2$ para a base $b = 16$

E 2.1.8. Quantos algarismos são necessários para representar o número 937163832173947 em base binária? E em base 7? Dica: Qual é o menor e o maior inteiro que pode ser escrito em dada base com N algarismos?

2.2 Notação científica e notação normalizada

Como vimos, no sistema posicional usual um número x na base b é representado por

$$x = \pm(d_n d_{n-1} \cdots d_0{,}d_{-1}d_{-2}d_{-3}\cdots)_b, \tag{2.26}$$

onde $d_n \neq 0$ e $d_i \in \{0, 1, \ldots, b-1\}$ é o dígito da i-ésima posição. Alternativamente, é costumeiro usarmos a chamada notação científica. Nesta, o número x é representado como

$$x = \pm(M)_b \times b^e, \tag{2.27}$$

onde $(M)_b = (d_m d_{m-1} \cdots d_0{,}d_{-1}d_{-2}d_{-3}\cdots)_b$ é chamada de mantissa e $e \in \mathbb{Z}$ é chamado de expoente de x.

Exemplo 2.2.1. a) O número 602,2141 em notação científica pode ser escrito como

$$602{,}2141 \times 10^0 = 60{,}22141 \times 10^1 = 0{,}6022141 \times 10^3. \tag{2.28}$$

b) O número $(1010{,}10)_2$ pode ser escrito em notação científica como $(10{,}1010)_2 \times 2^2$.

Observamos que um número pode ser representado de várias formas equivalentes em notação científica. Para termos uma representação única introduzimos o conceito de notação normalizada.

Definição 2.2.1. *Um número x na base b é dito estar representado em notação (científica) normalizada quando está escrito na forma*

$$x = (-1)^s (M)_b \times b^E, \tag{2.29}$$

onde $(M)_b = (d_0{,}d_{-1}d_{-2}d_{-3}\cdots)_b$, com $d_0 \neq 0$[3][4], s é 0 para positivo e 1 para negativo, E é o expoente.

Exemplo 2.2.2. Vejamos os seguintes casos:

a) O número 602,2141 em notação (científica) normalizada é representado por $6{,}022141 \times 10^2$.

b) O número $(1010{,}10)_2$ escrito em notação normalizada é $(1{,}01010)_2 \times 2^3$.

Observação 2.2.1. No `GNU Octave`, podemos controlar a impressão de números usando o comando `printf`. Por exemplo:

[3]Em algumas referências é usado $M_b = (0{,}d_{-1}d_{-2}d_{-3}\cdots)_b$.
[4]No caso de $x = 0$, $M_b = (0{,}00\cdots)_b$.

```
>> printf('%1.5f\n',-pi)
-3.14159
>> printf('%1.5e\n',-pi)
-3.14159e+00
```

No primeiro caso, obtemos a representação em ponto flutuante decimal com 6 dígitos do número $-\pi$. No segundo caso, obtemos a representação em notação científica normalizada com 6 dígitos.

Exercícios resolvidos

Exercícios

Esta seção carece de exercícios. Participe da sua escrita.

Veja como em:

https://github.com/livroscolaborativos/CalculoNumerico

E 2.2.1. Represente os seguintes números em notação científica normalizada:

$$\begin{array}{ll} a)\ 299792{,}458 & b)\ 66{,}2607 \times 10^{-35} \\ c)\ 0{,}6674 \times 10^{-7} & d)\ 9806{,}65 \times 10^{1} \end{array} \tag{2.30}$$

(2.31)

E 2.2.2. Use o computador para verificar as respostas do Exercício 2.2.1.

2.3 Representação decimal finita

Em computadores, é usual representarmos números usando uma quantidade de dígitos finita. A quantidade a ser usada normalmente depende da precisão com que as computações estão sendo feitas. Ocorre que quando restringimos a representação a um número finito de dígitos, muitos números não podem ser representado de forma exata, por exemplo, as dízimas infinitas e os números irracionais. Este fenômeno nos leva aos conceitos de número de dígitos significativos e de arredondamento.

Definição 2.3.1 (Número de dígitos significativos)**.** *Um número decimal* $x = \pm d_0{,}d_{-1}\cdots d_{-i}d_{-i-1}\cdots d_{-i-n}d_{-i-n-1}\cdot\times 10^E$ *é dito ter* n *dígitos significativos quando* $d_j = 0$ *para* $j \geq -i$ *e* $j \leq -i-n-1$.

Exemplo 2.3.1. O número $0{,}0602100 \times 10^{-3}$ tem 4 dígitos significativos.

2.3.1 Arredondamento de números

Quando representamos um número x com uma quantidade de dígitos menor que a de dígitos significativos acabamos com uma aproximação deste. Este procedimento é chamado arredondamento de um número. Mais precisamente, seja dado

$$x = \pm d_0{,}d_1d_2\ldots d_{k-1}d_kd_{k+1}\ldots d_n \times 10^e \tag{2.33}$$

em notação normalizada, isto é, $d_0 \neq 0$[5]. Podemos representar x com k dígitos da seguinte forma:

1. **Arredondamento por truncamento** (ou corte): aproximamos x por

$$\bar{x} = \pm d_0{,}d_1d_2\ldots d_k \times 10^e \tag{2.34}$$

simplesmente descartando os dígitos d_j com $j > k$.

2. **Arredondamento por proximidade**[6]: se $d_{k+1} < 5$ aproximamos x por

$$\bar{x} = \pm d_0{,}d_1d_2\ldots d_k \times 10^e \tag{2.35}$$

senão aproximamos x por[7]

$$\bar{x} = \pm(d_0{,}d_1d_2\ldots d_k + 10^{-k}) \times 10^e \tag{2.37}$$

3. **Arredondamento por proximidade com desempate par**: se $d_{k+1} < 5$ aproximamos x por

$$\bar{x} = \pm d_0{,}d_1d_2\ldots d_k \times 10^e. \tag{2.38}$$

Se $d_{k+1}{,}d_{d+2}d_{k+3}\ldots > 5$ aproximamos x por

$$\bar{x} = \pm(d_0{,}d_1d_2\ldots d_k + 10^{-k}) \times 10^e. \tag{2.39}$$

Agora, no caso de empate, i.e. $d_{k+1}{,}d_{k+2}d_{k+3}\ldots = 5$, então x é aproximado por

$$\bar{x} = \pm d_0{,}d_1d_2\ldots d_k \times 10^e \tag{2.40}$$

caso d_k seja par e, caso contrário, por

$$\bar{x} = \pm(d_0{,}d_1d_2\ldots d_k + 10^{-k}) \times 10^e. \tag{2.41}$$

[5] caso $x \neq 0$.

[6] com desempate infinito.

[7] Note que essas duas opções são equivalentes a somar 5 no dígito à direita do corte e depois arredondar por corte, ou seja, arredondar por corte

$$\pm(d_0{,}d_1d_2\ldots d_kd_{k+1} + 5 \times 10^{-(k+1)}) \times 10^e \tag{2.36}$$

Observação 2.3.1. O arredondamento por proximidade é frequentemente empregado para arredondamentos de números reais para inteiros. Por exemplo:

- $x = 1{,}49$ arredonda-se para o inteiro 1.
- $x = 1{,}50$ arredonda-se para o inteiro 2.
- $x = 2{,}50$ arredonda-se para o inteiro 3.

```
>> round(1.49)
ans =  1
>> round(1.50)
ans =  2
>> round(2.50)
ans =  3
```

Exemplo 2.3.2. Represente os números $x_1 = 0{,}567$, $x_2 = 0{,}233$, $x_3 = -0{,}675$ e $x_4 = 0{,}314159265\ldots \times 10^1$ com dois dígitos significativos por truncamento e arredondamento.

Solução. Vejamos cada caso:

- Por truncamento:

$$x_1 = 0{,}56, \quad x_2 = 0{,}23, \quad x_3 = -0{,}67 \quad \text{e} \quad x_4 = 3{,}1. \tag{2.42}$$

No `GNU Octave`, podemos obter a representação de $x_3 = -0{,}675$ fazendo:

```
>> printf("%1.2f\n", ceil(-0.675*1e2)/1e2)
-0.67
```

e, em notação normalizada, temos:

```
>> printf("%1.1e\n", ceil(-0.675*1e2)/1e2)
-6.7e-01
```

Em `GNU Octave`, a representação de números por arredondamento por proximidade com desempate par é o padrão. Assim, para obtermos a representação desejada de $x_3 = 0{,}675$ fazemos:

```
>> printf("%1.2f\n", -0.675)
-0.68
```

e, em notação normalizada, temos:

```
>> printf("%1.1e\n", -0.675)
-6.8e-01
```

$\Diamond$

Observação 2.3.2. Observe que o arredondamento pode mudar todos os dígitos e o expoente da representação em ponto flutuante de um número dado. Por exemplo, o arredondamento de $0{,}9999 \times 10^{-1}$ com 3 dígitos significativos é $0{,}1 \times 10^{0}$.

Exercícios resolvidos

Exercícios

E 2.3.1. Aproxime os seguintes números para 2 dígitos significativos por arredondamento por truncamento.

(a) 1,159

(b) 7,399

(c) $-5{,}901$

E 2.3.2. Aproxime os seguintes números para 2 dígitos significativos por arredondamento por proximidade com desempate par.

(a) 1,151

(b) 1,15

(c) 2,45

(d) $-2{,}45$

E 2.3.3. O `GNU Octave` usa arredondamento por proximidade com desempate par como padrão. Assim sendo, por exemplo

```
>> printf('%1.1e\n', 1.25)
1.2e+00
```

Agora:

```
>> printf('%1.1e\n', 2.45)
2.5e+00
```

Não deveria ser 2.4? Explique o que está ocorrendo.

2.4 Representação de números em máquina

Os computadores, em geral, usam a base binária para representar os números, onde as posições, chamadas de bits, assumem as condições "verdadeiro" ou "falso", ou seja, 1 ou 0, respectivamente. Os computadores representam os números com uma quantidade fixa de bits, o que se traduz em um conjunto finito de números representáveis. Os demais números são tomados por proximidade àqueles conhecidos, gerando erros de arredondamento. Por exemplo, em aritmética de computador, o número 2 tem representação exata, logo $2^2 = 4$, mas $\sqrt{3}$ não tem representação finita, logo $(\sqrt{3})^2 \neq 3$.

Veja isso no `GNU Octave`:

```
>> 2^2 == 4
ans =  1
>> sqrt(3)^2 == 3
ans = 0
```

2.4.1 Números inteiros

Tipicamente, um número inteiro é armazenado em um computador como uma sequência de dígitos binários de comprimento fixo denominado **registro**.

Representação sem sinal

Um registro com n bits da forma

d_{n-1}	d_{n-2}	$\cdots$	d_1	d_0

representa o número $(d_{n-1}d_{n-2}...d_1d_0)_2$.

Assim, é possível representar números inteiros entre $2^n - 1$ e 0, sendo

$$\begin{aligned} [111\ldots111] &= 2^{n-1} + 2^{n-2} + \cdots + 2^1 + 2^0 = 2^n - 1, \\ &\vdots \\ [000\ldots011] &= 3, \\ [000\ldots010] &= 2, \\ [000\ldots001] &= 1, \\ [000\ldots000] &= 0. \end{aligned} \tag{2.43}$$

Exemplo 2.4.1. No `Scilab`,

```
-->uint8( bin2dec('00000011') )
   ans = 3
-->uint8( bin2dec('11111110') )
   ans = 254
```

Representação com bit de sinal

O bit mais significativo (o primeiro à esquerda) representa o sinal: por convenção, 0 significa positivo e 1 significa negativo. Um registro com n bits da forma

s	d_{n-2}	$\cdots$	d_1	d_0

representa o número $(-1)^s(d_{n-2}\dots d_1d_0)_2$. Assim, é possível representar números inteiros entre $1-2^{n-1}$ e $2^{n-1}-1$, com duas representações para o zero: $(1000\dots000)_2$ e $(00000\dots000)_2$.

Exemplo 2.4.2. Em um registro com 8 bits, teremos os números

$$\begin{aligned}
[11111111] &= -(2^6+\cdots+2+1) = -127,\\
&\vdots\\
[10000001] &= -1,\\
[10000000] &= -0,\\
[01111111] &= 2^6+\cdots+2+1 = 127,\\
&\vdots\\
[00000010] &= 2,\\
[00000001] &= 1,\\
[00000000] &= 0.
\end{aligned} \tag{2.44}$$

Representação complemento de dois

O bit mais significativo (o primeiro à esquerda) representa o coeficiente de -2^{n-1}. Um registro com n bits da forma:

d_{n-1}	d_{n-2}	$\cdots$	d_1	d_0

representa o número $-d_{n-1}2^{n-1}+(d_{n-2}\dots d_1d_0)_2$.

Observação 2.4.1. Note que todo registro começando com 1 será um número negativo.

Exemplo 2.4.3. O registro com 8 bits [01000011] representa o número:

$$-0(2^7)+(1000011)_2 = 2^6+2+1 = 67. \tag{2.45}$$

Agora, o registro [10111101] representa:

$$-1(2^7)+(0111101)_2 = -128+2^5+2^4+2^3+2^2+1 = -67. \tag{2.46}$$

Note que podemos obter a representação de -67 invertendo os dígitos de 67 em binário e somando 1.

Exemplo 2.4.4. Em um registro com 8 bits, teremos os números

$$\begin{aligned}
&[11111111] = -2^7 + 2^6 + \cdots + 2 + 1 = -1\\
&\qquad\vdots\\
&[10000001] = -2^7 + 1 = -127\\
&[10000000] = -2^7 = -128\\
&[01111111] = 2^6 + \cdots + 2 + 1 = 127\\
&\qquad\vdots\\
&[00000010] = 2\\
&[00000001] = 1\\
&[00000000] = 0
\end{aligned} \tag{2.47}$$

O `GNU Octave` trabalha com representação complemento de 2 de números inteiros. O comando `bitunpack` mostra o barramento de um número dado, por exemplo:

```
>> bitunpack(int8(3))
ans =
   1   1   0   0   0   0   0   0
```

mostra o barramento com 8 **bits** do número inteiro 3. Note que a ordem dos **bits** é inversa daquela apresentada no texto acima. Aqui, o **bit** mais à esquerda fornece o coeficiente de 2^0, enquanto o **bit** mais à direita fornece o coeficiente de -2^7.

O comando `bitpack` converte um barramento para o número em decimal, por exemplo:

```
>> bitpack(logical([1 1 0 0 0 0 0 0]),'int8')
ans = 3
```

2.4.2 Sistema de ponto fixo

O sistema de ponto fixo representa as partes inteira e fracionária do número com uma quantidade fixa de dígitos.

Exemplo 2.4.5. Em um computador de 32 bits que usa o sistema de ponto fixo, o registro

d_{31}	d_{30}	d_{29}	$\cdots$	d_1	d_0

pode representar o número

- $(-1)^{d_{31}}(d_{30}d_{29}\cdots d_{17}d_{16}, d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal for representado por um dígito. Observe que, neste caso, o zero possui duas representações possíveis:

$$[10000000000000000000000000000000] \tag{2.48}$$

e

$$[00000000000000000000000000000000] \tag{2.49}$$

- $(d_{30}d_{29}\cdots d_{17}d_{16})_2 - d_{31}(2^{15} - 2^{-16}) + (0,d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal do número estiver representado por uma implementação em complemento de um. Observe que o zero também possui duas representações possíveis:

$$[11111111111111111111111111111111] \tag{2.50}$$

e

$$[00000000000000000000000000000000] \tag{2.51}$$

- $(d_{30}d_{29}\cdots d_{17}d_{16})_2 - d_{31}2^{15} + (0,d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal do número estiver representado por uma implementação em complemento de dois. Nesse caso o zero é unicamente representado por

$$[00000000000000000000000000000000] \tag{2.52}$$

Observe que 16 dígitos são usados para representar a parte fracionária, 15 são para representar a parte inteira e um dígito, o d_{31}, está relacionado ao sinal do número.

2.4.3 Sistema de ponto flutuante

O sistema de ponto flutuante não possui quantidade fixa de dígitos para as partes inteira e fracionária do número.

Podemos definir uma máquina F em ponto flutuante de dois modos:

$$F(\beta,|M|,|E|,BIAS) \text{ ou } F(\beta,|M|,E_{MIN},E_{MAX}) \tag{2.53}$$

onde

- β é a base (em geral 2 ou 10),
- $|M|$ é o número de dígitos da mantissa,
- $|E|$ é o número de dígitos do expoente,
- $BIAS$ é um valor de deslocamento do expoente (veja a seguir),

- E_{MIN} é o menor expoente,
- E_{MAX} é o maior expoente.

Considere uma máquina com um registro de 64 bits e base $\beta = 2$. Pelo padrão IEEE754, 1 bit é usado para o sinal, 11 bits para o expoente e 52 bits são usados para o significando tal que

s	c_{10}	c_9	$\cdots$	c_0	m_1	m_2	$\cdots$	m_{51}	m_{52}

represente o número (o $BIAS = 1023$ por definição)

$$x = (-1)^s M \times 2^{c-BIAS}, \tag{2.54}$$

onde a **característica** é representada por

$$c = (c_{10}c_9 \cdots c_1c_0)_2 = c_{10}2^{10} + \cdots + c_1 2^1 + c_0 2^0 \tag{2.55}$$

e o significando por

$$M = (1.m_1m_2 \cdots m_{51}m_{52})_2. \tag{2.56}$$

Observação 2.4.2. Em base 2 não é necessário armazenar o primeiro dígito (por quê?).

Exemplo 2.4.6. O registro

$$[0|100\ 0000\ 0000|1010\ 0000\ 0000 \ldots 0000\ 0000] \tag{2.57}$$

representa o número

$$(-1)^0(1 + 2^{-1} + 2^{-3}) \times 2^{1024-1023} = (1 + 0.5 + 0.125)2 = 3.25. \tag{2.58}$$

Observação 2.4.3. No `GNU Octave`, podemos usar os comandos `bitpack` e `bitunpack` transformar um registro de ponto flutuante de 64 **bits** em decimal e vice-versa. Entretanto, um tal registro no `GNU Octave` tem a seguinte estrutura

$$[m_{52}m_{51}m_{50} \ldots m_1|c_0c_1c_2 \cdots c_{10}|s]. \tag{2.59}$$

Desta forma, o decimal 3.25 tem, aqui, o registro

$$[000 \ldots 0101|000 \ldots 01|0]. \tag{2.60}$$

O que podemos verificar com o comando

```
>> bitpack(logical([zeros(1,49) 1 0 1 zeros(1,10) 1 0]),'double')
ans =  3.2500
```

O expoente deslocado

Uma maneira de representar os expoentes inteiros é deslocar todos eles uma mesma quantidade. Desta forma permitimos a representação de números negativos e a ordem deles continua crescente. O expoente é representado por um inteiro sem sinal do qual é deslocado o **BIAS**.

Tendo $|E|$ dígitos para representar o expoente, geralmente o $BIAS$ é predefinido de tal forma a dividir a tabela ao meio de tal forma que o expoente *um* seja representado pelo sequência $[100\ldots000]$.

Exemplo 2.4.7. Com 64 bits, pelo padrão $IEEE754$, temos que $|E| := 11$. Assim, $(100\ 0000\ 0000)_2 = 2^{10} = 1024$. Como queremos que esta sequência represente o 1, definimos $BIAS := 1023$, pois

$$1024 - BIAS = 1. \tag{2.61}$$

Com 32 bits, temos $|E| := 8$ e $BIAS := 127$. E com 128 bits, temos $|E| := 15$ e $BIAS := 16383$.

Com $|E| = 11$ temos

$$\begin{aligned}
&[111\ 1111\ 1111] = \text{reservado}\\
&[111\ 1111\ 1110] = 2046 - BIAS = 1023_{10} = E_{MAX}\\
&\qquad\qquad\vdots =\\
&[100\ 0000\ 0001] = 2^{10} + 1 - BIAS = 2_{10}\\
&[100\ 0000\ 0000] = 2^{10} \quad - BIAS = 1_{10}\\
&[011\ 1111\ 1111] = 1023 \quad - BIAS = 0_{10}\\
&[011\ 1111\ 1110] = 1022 \quad - BIAS = -1_{10}\\
&\qquad\qquad\vdots =\\
&[000\ 0000\ 0001] = 1 - BIAS = -1022 = E_{MIN}\\
&[000\ 0000\ 0000] = \text{reservado}
\end{aligned} \tag{2.62}$$

O maior expoente é dado por $E_{MAX} = 1023$ e o menor expoente é dado por $E_{MIN} = -1022$.

O menor número representável positivo é dado pelo registro

$$[0|000\ 0000\ 0001|0000\ 0000\ 0000\ldots0000\ 0000] \tag{2.63}$$

quando $s = 0$, $c = 1$ e $M = (1.000...000)_2$, ou seja,

$$MINR = (1+0)_2 \times 2^{1-1023} \approx 0.2225 \times 10^{-307}. \tag{2.64}$$

O maior número representável é dado por

$$[0|111\ 1111\ 1110|1111\ 1111\ \cdots\ 1111\ 1111] \tag{2.65}$$

quando $s = 0$, $c = 2046$ e $M = (1.1111\ 1111 \cdots\ 1111)_2 = 2 - 2^{-52}$, ou seja,

$$MAXR = (2 - 2^{-52}) \times 2^{2046-1023} \approx 2^{1024} \approx 0.17977 \times 10^{309}. \tag{2.66}$$

Observação 2.4.4. No `GNU Octave`, podemos obter o maior e o menor `double` positivo não nulo com os comandos:

```
>> realmax
ans =   1.7977e+308
>> realmin
ans =   2.2251e-308
```

Casos especiais

O **zero** é um caso especial representado pelo registro

$$[0|000\ 0000\ 0000|0000\ 0000\ 0000...0000\ 0000] \tag{2.67}$$

Os expoentes **reservados** são usados para casos especiais:

- $c = [0000...0000]$ é usado para representar o zero (se $m = 0$) e os números subnormais (se $m \neq 0$).
- $c = [1111...1111]$ é usado para representar o infinito (se $m = 0$) e NaN (se $m \neq 0$).

Os números subnormais[8] tem a forma

$$x = (-1)^s(0.m_1m_2 \cdots m_{51}m_{52})_2 \times 2^{1-BIAS}. \tag{2.68}$$

2.4.4 Precisão e épsilon de máquina

A **precisão** p de uma máquina é o número de dígitos significativos usado para representar um número. Note que $p = |M| + 1$ em binário e $p = |M|$ para outras bases.

O **épsilon de máquina**, $\epsilon_{mach} = \epsilon$, é definido de forma que $1 + \epsilon$ seja o menor número representável maior que 1, isto é, $1 + \epsilon$ é representável, mas não existem números representáveis em $(1, 1 + \epsilon)$.

[8] Note que poderíamos definir números um pouco menores que o $MINR$.

Exemplo 2.4.8. Com 64 bits, temos que o épsilon será dado por

$$\begin{aligned} 1 \rightarrow \quad & (1.0000\ 0000....0000)_2 \times 2^0 \\ \epsilon \rightarrow + & (0.0000\ 0000....0001)_2 \times 2^0 = 2^{-52} \\ & (1.0000\ 0000....0001)_2 \times 2^0 \neq 1 \end{aligned} \tag{2.69}$$

Assim, $\epsilon = 2^{-52}$.

Observação 2.4.5. No `GNU Octave`, o épsilon de máquina é representado pela constante `eps`. Observe os seguintes resultados:

```
>> 1 + 1e-16 == 1
ans =  1
>> 1 + eps == 1
ans = 0
```

2.4.5 Distribuição dos números

Utilizando uma máquina em ponto flutuante, temos um número finito de números que podemos representar.

Um número muito pequeno geralmente é aproximado por zero (**underflow**) e um número muito grande (**overflow**) geralmente faz o cálculo parar. Além disso, os números não estão uniformemente espaçados no eixo real. Números pequenos estão bem próximos enquanto que números com expoentes grandes estão bem distantes.

Se tentarmos armazenar um número que não é representável, devemos utilizar o número mais próximo, gerando os erros de arredondamento.

Observação 2.4.6. Veja como o `GNU Octave` se comporta nos seguintes casos de exceção:

```
>> 0/0
warning: division by zero
ans = NaN
>> 1/0
warning: division by zero
ans = Inf
>> 1/-0
warning: division by zero
ans = -Inf
>> 2*realmax
ans = Inf
>> 1/2^9999
ans = 0
```

Exercícios

E 2.4.1. Usando a representação complemento de dois de números inteiros com 8 **bits**, escreva o número decimal que corresponde aos seguintes barramentos:

a) [01100010].

b) [00011101].

c) [10000000].

d) [11100011].

e) [11111111]

E 2.4.2. Usando a representação complemento de dois de números inteiros com 16 **bits**, escreva o número decimal que corresponde aos seguintes barramentos:

a) [0110001001100010].

b) [0001110100011101].

c) [1110001011100011].

d) [1111111111111111].

E 2.4.3. Usando a representação complemento de dois de números inteiros com 8 **bits** no `GNU Octave`, escreva o número decimal que corresponde aos seguintes barramentos:

a) [01100010].

b) [00011101].

c) [00010010].

E 2.4.4. Usando a representação complemento de dois de números inteiros com 16 **bits** no `GNU Octave`, escreva o número decimal que corresponde aos seguintes barramentos:

a) [0110001001100010].

b) [0001110100011101].

c) [0001001011100010].

E 2.4.5. Usando a representação de ponto flutuante com 64 **bits**, escreva o número decimal que corresponde aos seguintes barramentos:

a) [0|10000000000|111000 . . . 0].

b) [1|10000000001|0111000 . . . 0].

E 2.4.6. Explique a diferença entre o sistema de ponto fixo e ponto flutuante.

E 2.4.7. Usando a representação de `double` no `GNU Octave`, escreva o número decimal que corresponde aos seguintes barramentos:

a) [000 . . . 0111|00000000001|0].

b) [000 . . . 01110|10000000001|1].

E 2.4.8. Considere a seguinte rotina escrita para ser usada no `GNU Octave`:

```
x=1
while x+1>x
    x=x+1
end
```

Explique se esta rotina finaliza em tempo finito, em caso afirmativo calcule a ordem de grandeza do tempo de execução supondo que cada passo do laço demore $10^{-7}s$. Justifique sua reposta.

2.5 Tipos de erros

Em geral, os números não são representados de forma exata nos computadores. Isto nos leva ao chamado erro de arredondamento. Quando resolvemos problemas com técnicas numéricas, estamos sujeitos a este e outros tipos de erros. Nesta seção, veremos quais são estes erros e como controlá-los, quando possível.

Quando fazemos aproximações numéricas, os erros são gerados de várias formas, sendo as principais delas as seguintes:

1. **Incerteza dos dados** são devidos aos erros nos dados de entrada. Quando o modelo matemático é oriundo de um problema físico, existe incerteza nas medidas feitas pelos instrumentos de medição, que possuem acurácia finita.

2. **Erros de Arredondamento** são aqueles relacionados com as limitações existentes na forma de representar números em máquina.

3. **Erros de Truncamento** surgem quando aproximamos um conceito matemático formado por uma sequência infinita de passos por de um procedimento finito. Por exemplo, a definição de integral é dada por um processo de limite de somas. Numericamente, aproximamos por um soma finita. O erro de truncamento deve ser estudado analiticamente para cada método empregado e normalmente envolve matemática mais avançada que a estudado em um curso de graduação.

Uma questão fundamental é a quantificação dos erros imbricados na computação da solução de um dado problema. Para tanto, precisamos definir medidas de erros (ou de exatidão). As medidas de erro mais utilizadas são o **erro absoluto** e o **erro relativo**.

Definição 2.5.1 (Erro absoluto e relativo)**.** *Seja x um número real e $\overline{x}$, sua aproximação. O* ***erro absoluto*** *da aproximação $\overline{x}$ é definido como*

$$|x - \overline{x}|. \tag{2.70}$$

O ***erro relativo*** *da aproximação $\overline{x}$ é definido como*

$$\frac{|x - \overline{x}|}{|x|}, \quad x \neq 0. \tag{2.71}$$

Observação 2.5.1. Observe que o erro relativo é adimensional e, muitas vezes, é expresso em porcentagens. Mais precisamente, o erro relativo em porcentagem da aproximação $\overline{x}$ é dado por

$$\frac{|x - \bar{x}|}{|x|} \times 100\%. \tag{2.72}$$

Exemplo 2.5.1. Sejam $x = 123456{,}789$ e sua aproximação $\bar{x} = 123000$. O erro absoluto é

$$|x - \bar{x}| = |123456{,}789 - 123000| = 456{,}789 \tag{2.73}$$

e o erro relativo é

$$\frac{|x - \bar{x}|}{|x|} = \frac{456{,}789}{123456{,}789} \approx 0{,}00369999 \text{ ou } 0{,}36\% \tag{2.74}$$

Exemplo 2.5.2. Sejam $y = 1{,}23456789$ e $\bar{y} = 1{,}13$. O erro absoluto é

$$|y - \bar{y}| = |1{,}23456789 - 1{,}13| = 0{,}10456789 \tag{2.75}$$

que parece pequeno se compararmos com o exemplo anterior. Entretanto o erro relativo é

$$\frac{|y-\bar{y}|}{|y|} = \frac{0{,}10456789}{1{,}23456789} \approx 0{,}08469999 \text{ ou } 8{,}4\% \tag{2.76}$$

Note que o erro relativo leva em consideração a escala do problema.

Exemplo 2.5.3. Observe os erros absolutos e relativos em cada caso a seguir:

x	$\bar{x}$	Erro absoluto	Erro relativo
$0{,}\overline{3} \times 10^{-2}$	$0{,}3 \times 10^{-2}$	$0{,}\overline{3} \times 10^{-3}$	10%
$0{,}\overline{3}$	$0{,}3$	$0{,}\overline{3} \times 10^{-2}$	10%
$0{,}\overline{3} \times 10^{2}$	$0{,}3 \times 10^{2}$	$0{,}\overline{3} \times 10^{1}$	10%

Outra forma de medir a exatidão de uma aproximação numérica é contar o **número de dígitos significativos corretos** em relação ao valor exato.

Definição 2.5.2 (Número de dígitos significativos corretos)**.** *A aproximação $\overline{x}$ de um número x tem s* ***dígitos significativos corretos*** *quando*[9]

$$\frac{|x-\overline{x}|}{|x|} < 5 \times 10^{-s}. \tag{2.78}$$

Exemplo 2.5.4. Vejamos os seguintes casos:

a) A aproximação de $x = 0{,}333333$ por $\overline{x} = 0{,}333$ tem 3 dígitos significativos corretos, pois

$$\frac{|x-\overline{x}|}{|x|} = \frac{0{,}000333}{0{,}333333} \approx 0{,}000999 \leq 5 \times 10^{-\mathbf{3}}. \tag{2.79}$$

b) Considere as aproximações $\bar{x}_1 = 0{,}666$ e $\bar{x}_2 = 0{,}667$ de $x = 0{,}666888$. Os erros relativos são

$$\frac{|x-\bar{x}_1|}{|x|} = \frac{|0{,}666888 - 0{,}666|}{0{,}666888} \approx 0{,}00133... < 5 \times 10^{-3}. \tag{2.80}$$

[9]Esta definição é apresentada em [**?**]. Não existe uma definição única na literatura para o conceito de dígitos significativos corretos, embora não precisamente equivalentes, elas transmitem o mesmo conceito. Uma maneira de interpretar essa regra é: calcula-se o erro relativo na forma normalizada e a partir da ordem do expoente temos o número de dígitos significativos corretos. Como queremos o expoente, podemos estimar s por

$$DIGSE(x,\bar{x}) = s \approx int\left|\log_{10}\frac{|x-\overline{x}|}{|x|}\right|. \tag{2.77}$$

$$\frac{|x-\bar{x}_2|}{|x|} = \frac{|0{,}666888 - 0{,}667|}{0{,}666888} \approx 0{,}000167... < 5 \times 10^{-4}. \tag{2.81}$$

Note que $\bar{x}_1$ possui 3 dígitos significativos corretos e $\bar{x}_2$ possui 4 dígitos significativos (o quarto dígito é o dígito 0 que não aparece à direita, i.e, $\bar{x}_2 = 0.6670$. Isto também leva a conclusão que x_2 aproxima melhor o valor de x do que x_1 pois está mais próximo de x.

c) $\overline{x} = 9{,}999$ aproxima $x = 10$ com 4 dígitos significativos corretos, pois

$$\frac{|x-\overline{x}|}{|x|} = \frac{|10 - 9{,}999|}{10} \approx 0{,}0000999... < 5 \times 10^{-4}. \tag{2.82}$$

d) Considere as aproximações $\overline{x}_1 = 1{,}49$ e $\overline{x}_2 = 1{,}5$ de $x = 1$. Da definição, temos que 1,49 aproxima 1 com um dígito significativo correto (verifique), enquanto 1,5 tem zero dígito significativo correto, pois:

$$\frac{|1 - 1{,}5|}{|1|} = 5 \times 10^{-1} < 5 \times 10^{0}. \tag{2.83}$$

Exercícios

E 2.5.1. Calcule os erros absoluto e relativo das aproximações $\bar{x}$ para x em cada caso:

a) $x = \pi = 3{,}14159265358979\ldots$ e $\bar{x} = 3{,}141$

b) $x = 1{,}00001$ e $\bar{x} = 1$

c) $x = 100001$ e $\bar{x} = 100000$

E 2.5.2. Arredonde os seguintes números para cinco algarismos significativos:

a) 1,7888544

b) 1788,8544

c) 0,0017888544

d) 0,004596632

e) $2{,}1754999 \times 10^{-10}$

f) $2{,}1754999 \times 10^{10}$

E 2.5.3. Represente os seguintes números com três dígitos significativos usando arredondamento por truncamento e arredondamento por proximidade.

a) 3276.

b) 42,55.

c) 0,00003331.

E 2.5.4. Usando a Definição 2.5.2, verifique quantos são os dígitos significativos corretos na aproximação de x por $\bar{x}$.

a) $x = 2{,}5834$ e $\bar{x} = 2{,}6$

b) $x = 100$ e $\bar{x} = 99$

E 2.5.5. Resolva a equação $0{,}1x - 0{,}01 = 12$ usando arredondamento com três dígitos significativos em cada passo e compare com o resultado exato.

E 2.5.6. Calcule o erro relativo e absoluto envolvido nas seguintes aproximações e expresse as respostas com três algarismos significativos corretos.

a) $x = 3{,}1415926535898$ e $\tilde{x} = 3{,}141593$

b) $x = \frac{1}{7}$ e $\tilde{x} = 1{,}43 \times 10^{-1}$

2.6 Erros nas operações elementares

O erro relativo presente nas operações elementares de adição, subtração, multiplicação e divisão é da ordem do épsilon de máquina. Se estivermos usando o sistema de numeração *binary64* ou *double*, temos $\epsilon = 2^{-52} \approx 2{,}22 \cdot 10^{-16}$.

Este erro é bem pequeno para a maioria das aplicações! Assumindo que x e y são representados com todos dígitos corretos, esperamos ter aproximadamente 15 dígitos significativos corretos quando fazemos uma das operações $x + y$, $x - y$, $x \times y$ ou x/y.

Mesmo que fizéssemos, por exemplo, 1000 operações elementares sucessivas em ponto flutuante, teríamos, no pior dos casos, acumulado todos esses erros e perdido 3 casas decimais ($1000 \times 10^{-15} \approx 10^{-12}$).

Entretanto, existem situações em que o erro se propaga de forma muito catastrófica, em especial, quando subtraímos números positivos muito próximos.

2.7 Cancelamento catastrófico

Quando fazemos subtrações com números muito próximos entre si, ocorre o que chamamos de "cancelamento catastrófico", onde podemos perder vários dígitos de precisão em uma única subtração.

Exemplo 2.7.1. Efetue a operação

$$0{,}987624687925 - 0{,}987624 = 0{,}687925 \times 10^{-6} \tag{2.87}$$

usando arredondamento com seis dígitos significativos e observe a diferença se comparado com resultado sem arredondamento.

Solução. Os números arredondados com seis dígitos para a mantissa resultam na seguinte diferença

$$0{,}987625 - 0{,}987624 = 0{,}100000 \times 10^{-5} \tag{2.88}$$

Observe que os erros relativos entre os números exatos e aproximados no lado esquerdo são bem pequenos,

$$\frac{|0{,}987624687925 - 0{,}987625|}{|0{,}987624687925|} = 0{,}00003159 \tag{2.89}$$

e

$$\frac{|0{,}987624 - 0{,}987624|}{|0{,}987624|} = 0\%, \tag{2.90}$$

enquanto no lado direito o erro relativo é enorme:

$$\frac{|0{,}100000 \times 10^{-5} - 0{,}687925 \times 10^{-6}|}{0{,}687925 \times 10^{-6}} = 45{,}36\%. \tag{2.91}$$

$\Diamond$

Exemplo 2.7.2. Considere o problema de encontrar as raízes da equação de segundo grau

$$x^2 + 300x - 0{,}014 = 0, \tag{2.92}$$

usando seis dígitos significativos.

Aplicando a fórmula de Bhaskara com $a = 0{,}100000 \times 10^1$, $b = 0{,}300000 \times 10^3$ e $c = 0{,}140000 \times 10^{-1}$, temos o discriminante:

$$\begin{aligned}
\Delta &= b^2 - 4 \cdot a \cdot c && (2.93)\\
&= 0{,}300000 \times 10^3 \times 0{,}300000 \times 10^3 && (2.94)\\
&+ 0{,}400000 \times 10^1 \times 0{,}100000 \times 10^1 \times 0{,}140000 \times 10^{-1} && (2.95)\\
&= 0{,}900000 \times 10^5 + 0{,}560000 \times 10^{-1} && (2.96)\\
&= 0{,}900001 \times 10^5 && (2.97)
\end{aligned}$$

e as raízes:

$$\begin{aligned} x_1,x_2 &= \frac{-0{,}300000 \times 10^3 \pm \sqrt{\Delta}}{0{,}200000 \times 10^1} && (2.98)\\ &= \frac{-0{,}300000 \times 10^3 \pm \sqrt{0{,}900001 \times 10^5}}{0{,}200000 \times 10^1} && (2.99)\\ &= \frac{-0{,}300000 \times 10^3 \pm 0{,}300000 \times 10^3}{0{,}200000 \times 10^1} && (2.100)\\ & && (2.101) \end{aligned}$$

Então, as duas raízes obtidas com erros de arredondamento, são:

$$\begin{aligned} \tilde{x}_1 &= \frac{-0{,}300000 \times 10^3 - 0{,}300000 \times 10^3}{0{,}200000 \times 10^1} \\ &= -\frac{0{,}600000 \times 10^3}{0{,}200000 \times 10^1} = -0{,}300000 \times 10^3 \end{aligned} \tag{2.102}$$

e

$$\tilde{x}_2 = \frac{-0{,}300000 \times 10^3 + 0{,}300000 \times 10^3}{0{,}200000 \times 10^1} = 0{,}000000 \times 10^0 \tag{2.103}$$

No entanto, os valores das raízes com seis dígitos significativos livres de erros de arredondamento, são:

$$x_1 = -0{,}300000 \times 10^3 \quad \text{e} \quad x_2 = 0{,}466667 \times 10^{-4}. \tag{2.104}$$

Observe que a primeira raiz apresenta seis dígitos significativos corretos, mas a segunda não possui nenhum dígito significativo correto.

Observe que isto acontece porque b^2 é muito maior que $4ac$, ou seja, $b \approx \sqrt{b^2 - 4ac}$, logo a diferença

$$-b + \sqrt{b^2 - 4ac} \tag{2.105}$$

estará próxima de zero. Uma maneira de evitar o cancelamento catastrófico é aplicar procedimentos analíticos na expressão para eliminar essa diferença. Um técnica padrão consiste usar uma expansão em série de Taylor em torno da origem, tal como:

$$\sqrt{1-x} = 1 - \frac{1}{2}x + O(x^2). \tag{2.106}$$

Substituindo esta aproximação na fórmula de Bhaskara, temos:

$$\begin{aligned} x &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} && (2.107)\\ &= \frac{-b \pm b\sqrt{1 - \frac{4ac}{b^2}}}{2a} && (2.108)\\ &\approx \frac{-b \pm b\left(1 - \frac{4ac}{2b^2}\right)}{2a} && (2.109)\\ & && (2.110) \end{aligned}$$

Observe que $\frac{4ac}{b^2}$ é um número pequeno e por isso a expansão faz sentido. Voltamos no exemplo anterior e calculamos as duas raízes com o nova expressão

$$\begin{aligned} \tilde{x}_1 &= \frac{-b - b + \frac{4ac}{2b}}{2a} = -\frac{b}{a} + \frac{c}{b} && (2.111)\\ &= -\frac{0{,}300000 \times 10^3}{0{,}100000 \times 10^1} - \frac{0{,}140000 \times 10^{-1}}{0{,}300000 \times 10^3} && (2.112)\\ &= -0{,}300000 \times 10^3 - 0{,}466667 \times 10^{-4} && (2.113)\\ &= -0{,}300000 \times 10^3 && (2.114) \end{aligned}$$

$$\begin{aligned} \tilde{x}_2 &= \frac{-b + b - \frac{4ac}{2b}}{2a} && (2.115)\\ &= -\frac{4ac}{4ab} && (2.116)\\ &= -\frac{c}{b} = -\frac{-0{,}140000 \times 10^{-1}}{0{,}300000 \times 10^3} = 0{,}466667 \times 10^{-4} && (2.117)\\ & && (2.118) \end{aligned}$$

Observe que o efeito catastrófico foi eliminado.

Observação 2.7.1. O cancelamento catastrófico também poderia ter sido evitado através do seguinte truque analítico:

$$\begin{aligned} x_2 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} && (2.119)\\ &= \frac{b^2 - (b^2 - 4ac)}{2a\left(-b - \sqrt{b^2 - 4ac}\right)} = \frac{4ac}{2a\left(-b - \sqrt{b^2 - 4ac}\right)} && (2.120)\\ &= -\frac{2c}{\left(b + \sqrt{b^2 - 4ac}\right)} && (2.121) \end{aligned}$$

2.8 Condicionamento de um problema

Nesta seção, utilizaremos a seguinte descrição abstrata para o conceito de "resolver um problema": dado um conjunto de dados de entrada, encontrar os dados de saída. Se denotamos pela variável x os dados de entrada e pela variável y os dados de saída, resolver o problema significa encontrar y dado x. Em termos matemáticos, a resolução de um problema é realizada pelo mapeamento $f : x \to y$, ou simplesmente $y = f(x)$.

É certo que, na maioria das aplicações, os dados de entrada do problema — isto é, x — não são conhecidos com total exatidão, devido a diversas fontes de erros, como incertezas na coleta dos dados e erros de arredondamento. O conceito de condicionamento está relacionado à forma como os erros nos dados de entrada influenciam os dados de saída.

Para fins de análise, denotaremos por x, os dados de entrada com precisão absoluta e por x^*, os dados com erro. Definiremos também a solução y^*, do problema com dados de entrada x^*, ou seja, $y^* = f(x^*)$.

Estamos interessados em saber se os erros cometidos na entrada $\Delta x = x^* - x$ influenciaram na saída do problema $\Delta y = y^* - y$. No caso mais simples, temos que $x \in \mathbb{R}$ e $y \in \mathbb{R}$. Assumindo que f seja diferenciável, a partir da série de Taylor

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x \tag{2.122}$$

obtemos (subtraindo $f(x)$ dos dois lados)

$$\Delta y = f(x + \Delta x) - f(x) \approx f'(x)\Delta x \tag{2.123}$$

Para relacionarmos os erros relativos, dividimos o lado esquerdo por y, o lado direito por $f(x) = y$ e obtemos

$$\frac{\Delta y}{y} \approx \frac{f'(x)}{f(x)} \frac{x\Delta x}{x} \tag{2.124}$$

sugerindo a definição de número de condicionamento de um problema.

Definição 2.8.1. *Seja f uma função diferenciável. O* ***número de condicionamento*** *de um problema é definido como*

$$\kappa_f(x) := \left| \frac{xf'(x)}{f(x)} \right| \tag{2.125}$$

e fornece uma estimativa de quanto os erros relativos na entrada $\left|\frac{\Delta x}{x}\right|$ serão amplificados na saída $\left|\frac{\Delta y}{y}\right|$.

De modo geral, quando f depende de várias variáveis, podemos obter

$$\delta_f = |f(x_1,x_2,...,x_n) - f(\bar{x}_1,\bar{x}_2,...,\bar{x}_n)| \approx \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(x_1,x_2,...,x_n)\right| \delta_{x_i} \tag{2.126}$$

Uma matriz de números de condicionamento também poderia ser obtida como em [**?**].

Exemplo 2.8.1. Considere o problema de calcular $\sqrt{x}$ em $x = 2$. Se usarmos $x^* = 1{,}999$, quanto será o erro relativo na saída? O erro relativo na entrada é

$$\left|\frac{\Delta x}{x}\right| = \left|\frac{2 - 1{,}999}{2}\right| = 0{,}0005 \tag{2.127}$$

O número de condicionamento do problema calcular a raiz é

$$\kappa_f(x) := \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\frac{1}{2\sqrt{x}}}{\sqrt{x}}\right| = \frac{1}{2} \tag{2.128}$$

Ou seja, os erros na entrada serão diminuídos pela metade. De fato, usando $y = \sqrt{2} = 1{,}4142136...$ e $y^* = \sqrt{1{,}999} = 1{,}41386...$, obtemos

$$\frac{\Delta y}{y} = \frac{\sqrt{2} - \sqrt{1{,}999}}{\sqrt{2}} \approx 0{,}000250031... \tag{2.129}$$

Exemplo 2.8.2. Considere a função $f(x) = \frac{10}{1-x^2}$ e $x^* = 0{,}9995$ com um erro absoluto na entrada de 0,0001.

Calculando $y^* = f(x^*)$ temos

$$y^* = \frac{10}{1 - (0{,}9995)^2} \approx 10002{,}500625157739705173 \tag{2.130}$$

Mas qual é a estimativa de erro nessa resposta? Quantos dígitos significativos temos nessa resposta?

Sabendo que $f'(x) = 20x/(1 - x^2)^2$, o número de condicionamento é

$$\kappa_f(x) := \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{2x^2}{1 - x^2}\right| \tag{2.131}$$

o que nos fornece para $x^* = 0{,}9995$,

$$\kappa_f(0{,}9995) \approx 1998{,}5 \tag{2.132}$$

Como o erro relativo na entrada é

$$\left|\frac{\Delta x}{x}\right| = \left|\frac{0{,}0001}{0{,}9995}\right| \approx 0{,}00010005... \tag{2.133}$$

temos que o erro na saída será aproximadamente

$$\left|\frac{\Delta y}{y}\right| \approx \kappa_f(x)\left|\frac{\Delta x}{x}\right| \approx 1998{,}5 \times 0{,}00010005... \approx 0{,}1999 \tag{2.134}$$

ou seja um erro relativo de aproximadamente 19,99%.

Note que se usarmos $x_1 = 0{,}9994$ e $x_2 = 0{,}9996$ (ambos no intervalo do erro absoluto da entrada) encontramos

$$y_1^* \approx 8335{,}83 \tag{2.135}$$
$$y_2^* \approx 12520{,}50 \tag{2.136}$$

confirmando a estimativa de 19,99%.

Exemplo 2.8.3. Seja $f(x) = x\exp(x)$. Calcule o erro absoluto ao calcular $f(x)$ sabendo que $x = 2 \pm 0{,}05$.

Solução. Temos que $x \approx 2$ com erro absoluto de $\delta_x = 0{,}05$. Neste caso, calculamos δ_f, isto é, o erro absoluto ao calcular $f(x)$, por:

$$\delta_f = |f'(x)|\delta_x. \tag{2.137}$$

Como $f'(x) = (1+x)e^x$, temos:

$$\delta_f = |(1+x)e^x| \cdot \delta_x \tag{2.138}$$
$$= |3e^2| \cdot 0{,}05 = 1{,}1084. \tag{2.139}$$

Portanto, o erro absoluto ao calcular $f(x)$ quando $x = 2 \pm 0{,}05$ é de 1,1084. $\Diamond$

Exemplo 2.8.4. Calcule o erro relativo ao medir $f(x,y) = \frac{x^2+1}{x^2}e^{2y}$ sabendo que $x \approx 3$ é conhecido com 10% de erro e $y \approx 2$ é conhecido com 3% de erro.

Solução. Calculamos as derivadas parciais de f:

$$\frac{\partial f}{\partial x} = \frac{2x^3 - (2x^3+2x)}{x^4}e^{2y} = -\frac{2e^{2y}}{x^3} \tag{2.140}$$

e

$$\frac{\partial f}{\partial y} = 2\frac{x^2+1}{x^2}e^{2y} \tag{2.141}$$

Calculamos o erro absoluto em termos do erro relativo:

$$\frac{\delta_x}{|x|} = 0{,}1 \Rightarrow \delta_x = 3 \cdot 0{,}1 = 0{,}3 \tag{2.142}$$

$$\frac{\delta_y}{|y|} = 0{,}03 \Rightarrow \delta_y = 2 \cdot 0{,}03 = 0{,}06 \tag{2.143}$$

Aplicando a expressão para estimar o erro em f temos

$$\delta_f = \left|\frac{\partial f}{\partial x}\right| \delta_x + \left|\frac{\partial f}{\partial y}\right| \delta_y \tag{2.144}$$

$$= \frac{2e^4}{27} \cdot 0{,}3 + 2\frac{9+1}{9}e^4 \cdot 0{,}06 = 8{,}493045557 \tag{2.145}$$

Portanto, o erro relativo ao calcular f é estimado por

$$\frac{\delta f}{|f|} = \frac{8{,}493045557}{\frac{9+1}{9}e^4} = 14\% \tag{2.146}$$

$\Diamond$

Exemplo 2.8.5. No exemplo anterior, reduza o erro relativo em x pela metade e calcule o erro relativo em f. Depois, repita o processo reduzindo o erro relativo em y pela metade.

Solução. Na primeira situação temos $x = 3$ com erro relativo de 5% e $\delta_x = 0{,}05 \cdot 3 = 0{,}15$. Calculamos $\delta_f = 7{,}886399450$ e o erro relativo em f de 13%. Na segunda situação, temos $y = 2$ com erro de 1,5% e $\delta_y = 2 \cdot 0{,}015 = 0{,}03$. Calculamos $\delta_f = 4{,}853168892$ e o erro relativo em f de 8%. Observe que mesma o erro relativo em x sendo maior, o erro em y é mais significante na função. $\Diamond$

Exemplo 2.8.6. Considere um triângulo retângulo onde a hipotenusa e um dos catetos são conhecidos a menos de um erro: hipotenusa $a = 3 \pm 0{,}01$ metros e cateto $b = 2 \pm 0{,}01$ metros. Calcule o erro absoluto ao calcular a área dessa triângulo.

Solução. Primeiro vamos encontrar a expressão para a área em função da hipotenusa a e um cateto b. A tamanho de segundo cateto c é dado pelo teorema de Pitágoras, $a^2 = b^2 + c^2$, ou seja, $c = \sqrt{a^2 - b^2}$. Portanto a área é

$$A = \frac{bc}{2} = \frac{b\sqrt{a^2 - b^2}}{2}. \tag{2.147}$$

Agora calculamos as derivadas

$$\frac{\partial A}{\partial a} = \frac{ab}{2\sqrt{a^2 - b^2}}, \tag{2.148}$$

$$\frac{\partial A}{\partial b} = \frac{\sqrt{a^2 - b^2}}{2} - \frac{b^2}{2\sqrt{a^2 - b^2}}, \tag{2.149}$$

e substituindo na estimativa para o erro δ_A em termos de $\delta_a = 0{,}01$ e $\delta_b = 0{,}01$:

$$\delta_A \approx \left|\frac{\partial A}{\partial a}\right|\delta_a + \left|\frac{\partial A}{\partial b}\right|\delta_b \tag{2.150}$$

$$\approx \frac{3\sqrt{5}}{5}\cdot 0{,}01 + \frac{\sqrt{5}}{10}\cdot 0{,}01 = 0{,}01565247584 \tag{2.151}$$

Em termos do erro relativo temos erro na hipotenusa de $\frac{0{,}01}{3} \approx 0{,}333\%$, erro no cateto de $\frac{0{,}01}{2} = 0{,}5\%$ e erro na área de

$$\frac{0{,}01565247584}{\frac{2\sqrt{3^2-2^2}}{2}} = 0{,}7\% \tag{2.152}$$

$\Diamond$

Exercícios

E 2.8.1. Considere que a variável $x \approx 2$ é conhecida com um erro relativo de 1% e a variável $y \approx 10$ com um erro relativo de 10%. Calcule o erro relativo associado a z quando:

$$z = \frac{y^4}{1+y^4}e^x. \tag{2.153}$$

Suponha que você precise conhecer o valor de z com um erro de 0,5%. Você propõe uma melhoria na medição da variável x ou y? Explique.

E 2.8.2. A corrente I em ampéres e a tensão V em volts em uma lâmpada se relacionam conforme a seguinte expressão:

$$I = \left(\frac{V}{V_0}\right)^{\alpha}, \tag{2.154}$$

onde α é um número entre 0 e 1 e V_0 é tensão nominal em volts. Sabendo que $V_0 = 220 \pm 3\%$ e $\alpha = -0{,}8 \pm 4\%$, calcule a corrente e o erro relativo associado quando a tensão vale $220 \pm 1\%$.
Obs:. Este problema pode ser resolvido de duas formas distintas: usando a expressão aproximada para a propagação de erro e inspecionando os valores máximos e mínimos que a expressão pode assumir. Pratique os dois métodos.**Dica:** lembre que $x^{\alpha} = e^{\alpha \ln(x)}$

2.9 Exemplos selecionados de cancelamento catastrófico

Exemplo 2.9.1. Considere o seguinte processo iterativo:

$$x^{(1)} = \frac{1}{3} \tag{2.155}$$

$$x^{(n+1)} = 4x^{(n)} - 1, \quad n = 1, 2, \ldots \tag{2.156}$$

Observe que $x^{(1)} = \frac{1}{3}$, $x^{(2)} = 4 \cdot \frac{1}{3} - 1 = \frac{1}{3}$, $x^{(3)} = \frac{1}{3}$, ou seja, temos uma sequência constante igual a $\frac{1}{3}$. No entanto, ao calcularmos no computador, usando o sistema de numeração `double`, a sequência obtida não é constante e, de fato, diverge.

Faça o teste no `GNU Octave`, colocando:

```
>> x = 1/3
```

e itere algumas vezes a linha de comando:

```
>> x = 4*x-1
```

Para compreender o que acontece, devemos levar em consideração que o número $\frac{1}{3} = 0{,}\overline{3}$ possui um representação infinita tanto na base decimal quanto na base binária. Logo, sua representação de máquina inclui um erro de arredondamento. Seja ϵ a diferença entre o valor exato de $\frac{1}{3}$ e sua representação de máquina, isto é, $\tilde{x}^{(1)} = \frac{1}{3} + \epsilon$. A sequência efetivamente calculada no computador é:

$$\tilde{x}^{(1)} = \frac{1}{3} + \epsilon \tag{2.157}$$

$$\tilde{x}^{(2)} = 4x^{(1)} - 1 = 4\left(\frac{1}{3} + \epsilon\right) - 1 = \frac{1}{3} + 4\epsilon \tag{2.158}$$

$$\tilde{x}^{(3)} = 4x^{(2)} - 1 = 4\left(\frac{1}{3} + 4\epsilon\right) - 1 = \frac{1}{3} + 4^2\epsilon \tag{2.159}$$

$$\vdots \tag{2.160}$$

$$\tilde{x}^{(n)} = \frac{1}{3} + 4^{(n-1)}\epsilon \tag{2.161}$$

Portanto o limite da sequência diverge,

$$\lim_{x \to \infty} |\tilde{x}^{(n)}| = \infty \tag{2.162}$$

Qual o número de condicionamento desse problema?

Exemplo 2.9.2. Observe a seguinte identidade

$$f(x) = \frac{(1 + x) - 1}{x} = 1 \tag{2.163}$$

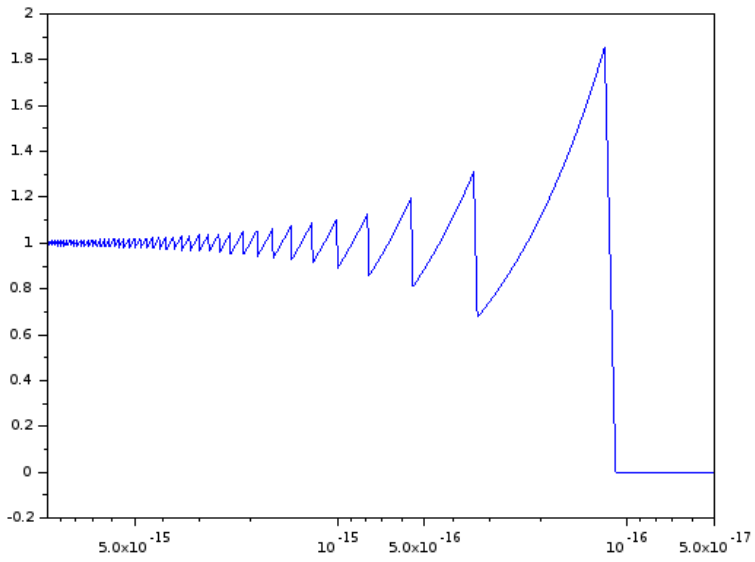


Figure 2.1: Gráfico na função do Exemplo 2.9.2.

Calcule o valor da expressão à esquerda para $x = 10^{-12}$, $x = 10^{-13}$, $x = 10^{-14}$, $x = 10^{-15}$, $x = 10^{-16}$ e $x = 10^{-17}$. Observe que quando x se aproxima do ϵ de máquina a expressão perde o significado. Veja a Figura 2.1 com o gráfico de $f(x)$ em escala logarítmica.

Exemplo 2.9.3. Neste exemplo, estamos interessados em compreender mais detalhadamente o comportamento da expressão

$$\left(1 + \frac{1}{n}\right)^n \tag{2.164}$$

quando n é um número grande ao computá-la em sistemas de numeral de ponto flutuante com acurácia finita. Um resultado bem conhecido do cálculo nos diz que o limite de (2.164) quando n tende a infinito é o número de Euler:

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e = 2{,}718281828459... \tag{2.165}$$

Sabemos também que a sequência produzida por (2.164) é crescente, isto é:

$$\left(1 + \frac{1}{1}\right)^1 < \left(1 + \frac{1}{2}\right)^2 < \left(1 + \frac{1}{3}\right)^3 < \cdots \tag{2.166}$$

No entanto, quando calculamos essa expressão no `Julia`, nos defrontamos com o seguinte resultado:

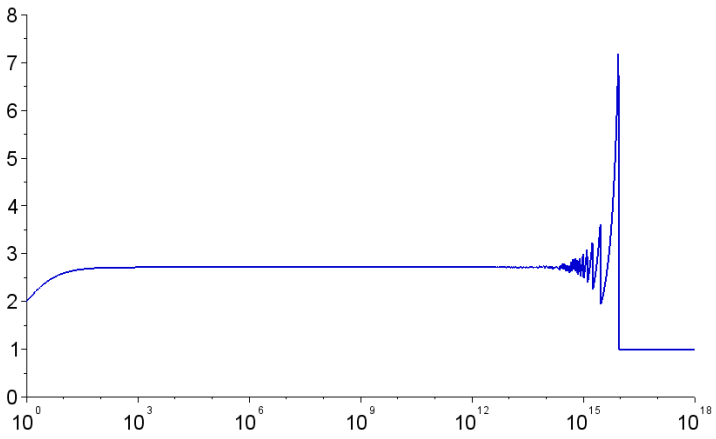


Figure 2.2: Gráfico de $\left(1+\frac{1}{n}\right)^n$ em função de n em escala linear-logarítmica variando de 10^0 até 10^{18}. Veja o Exemplo 2.9.3.

n	$\left(1+\frac{1}{n}\right)^n$		n	$\left(1+\frac{1}{n}\right)^n$
1	2,0000000000000		10^2	2,7048138294215
2	2,2500000000000		10^4	2,7181459268249
3	2,3703703703704		10^6	2,7182804690957
4	2,4414062500000		10^8	2,7182817983391
5	2,4883200000000		10^{10}	2,7182820532348
6	2,5216263717421		10^{12}	2,7185234960372
7	2,5464996970407		10^{14}	2,7161100340870
8	2,5657845139503		10^{16}	1,0000000000000
9	2,5811747917132		10^{18}	1,0000000000000
10	2,5937424601000		10^{20}	1,0000000000000

Podemos resumir esses dados no gráfico de $\left(1+\frac{1}{n}\right)^n$ em função de n, veja a Figura 2.9.

Observe que quando n se torna grande, da ordem de 10^{15}, o gráfico da função

deixa de ser crescente e apresenta oscilações. Observe também que a expressão se torna identicamente igual a 1 depois de um certo limiar. Tais fenômenos não são intrínsecos da função $f(n) = (1 + 1/n)^n$, mas **oriundas de erros de arredondamento**, isto é, são resultados numéricos espúrios. A fim de pôr o comportamento numérico de tal expressão, apresentamos abaixo o gráfico da mesma função, porém restrito à região entre 10^{14} e 10^{16}.

Para compreendermos melhor por que existe um limiar N que, quando atingido torna a expressão do exemplo acima identicamente igual a 1, observamos a sequência de operações realizadas pelo computador:

$$n \;\rightarrow\; 1/n \;\rightarrow\; 1 + 1/n \;\rightarrow\; (1 + 1/n)^n \tag{2.167}$$

Devido ao limite de precisão da representação de números em ponto flutuante, existe um menor número representável que é maior do que 1. Este número é 1+`eps`, onde `eps` é chamado de **épsilon de máquina** e é o menor número que somado a 1 produz um resultado superior a 1 no sistema de numeração usado. O épsilon de máquina no sistema de numeração **double** vale aproximadamente $2{,}22 \times 10^{-16}$.

No `GNU Octave`, o épsilon de máquina é a constante `eps`. Observe que:

```
>> printf('%1.25f\n', 1+eps)
1.0000000000000002220446049
```

Quando somamos a 1 um número positivo inferior ao épsilon de máquina, obtemos o número 1. Dessa forma, o resultado obtido pela operação de ponto flutuante $1 + n$ para $0 < n < 2{,}22 \times 10^{-16}$ é 1.

Portanto, quando realizamos a sequência de operações dada em (2.167), toda informação contida no número n é perdida na soma com 1 quando $1/n$ é menor que o épsilon de máquina, o que ocorre quando $n > 5 \times 10^{15}$. Assim, $(1 + 1/n)$ é aproximado para 1 e a última operação se resume a 1^n, o que é igual a 1 mesmo quando n é grande.

Um erro comum é acreditar que o perda de significância se deve ao fato de $1/n$ ser muito pequeno para ser representado e é aproximando para 0. Isto é falso, o sistema de ponto de flutuante permite representar números de magnitude muito inferior ao épsilon de máquina. O problema surge da limitação no tamanho da mantissa. Observe como a seguinte sequência de operações não perde significância para números positivos x muito menores que o épsilon de máquina:

$$n \;\rightarrow\; 1/n \;\rightarrow\; 1/(1/n) \tag{2.168}$$

compare o desempenho numérico desta sequência de operações para valores pequenos de n com o da seguinte sequência:

$$n \;\rightarrow\; 1 + n \;\rightarrow\; (1 + n) - 1. \tag{2.169}$$

Finalmente, notamos que quando tentamos calcular $\left(1+\frac{1}{n}\right)^n$ para n grande, existe perda de significância no cálculo de $1+1/n$.

Para entendermos isso melhor, vejamos o que acontece no `GNU Octave` quando $n = 7 \times 10^{13}$:

```
>> format('long')
>> n=7e13
n =  70000000000000
>> 1/n
ans =    1.42857142857143e-14
>> y=1+1/n
y =  1.00000000000001
```

Observe a perda de informação ao deslocar a mantissa de $1/n$. Para evidenciar o fenômenos, observamos o que acontece quando tentamos recalcular n subtraindo 1 de $1+1/n$ e invertendo o resultado:

```
>> y-1
ans =    1.42108547152020e-14
>> 1/(y-1)
ans =  70368744177664
```

Exemplo 2.9.4 (Analogia da balança)**.** Observe a seguinte comparação interessante que pode ser feita para ilustrar os sistemas de numeração com ponto fixo e flutuante: o sistema de ponto fixo é como uma balança cujas marcas estão igualmente espaçadas; o sistema de ponto flutuante é como uma balança cuja distância entre as marcas é proporcional à massa medida. Assim, podemos ter uma balança de ponto fixo cujas marcas estão sempre distanciadas de 100g (100g, 200g, 300g, ..., 1kg, 1,1kg,...) e outra balança de ponto flutuante cujas marcas estão distanciadas sempre de aproximadamente um décimo do valor lido (100g, 110g, 121g, 133g, ..., 1kg, 1,1kg, 1,21kg, ...) A balança de ponto fixo apresenta uma resolução baixa para pequenas medidas, porém uma resolução alta para grandes medidas. A balança de ponto flutuante distribui a resolução de forma proporcional ao longo da escala.

Seguindo nesta analogia, o fenômeno de perda de significância pode ser interpretado como a seguir: imagine que você deseje obter o peso de um gato (aproximadamente 4kg). Dois processos estão disponíveis: colocar o gato diretamente na balança ou medir seu peso com o gato e, depois, sem o gato. Na balança de ponto flutuante, a incerteza associada à medida do peso do gato (sozinho) é aproximadamente 10% de 4kg, isto é, 400g. Já a incerteza associada à medida da uma pessoa (aproximadamente 70kg) com o gato é de 10% do peso total, isto é, aproximadamente 7kg. Esta incerteza é da mesma ordem de grandeza da medida a ser realizada, tornado o processo impossível de ser realizado, já que teríamos uma incerteza da ordem de 14kg (devido à dupla medição) sobre uma grandeza de 4kg.

Exercícios resolvidos

ER 2.9.1. Deseja-se medir a concentração de dois diferentes oxidantes no ar. Três sensores eletroquímicos estão disponíveis para a medida e apresentam a seguintes respostas:

$$v_1 = 270[A] + 30[B], \quad v_2 = 140[A] + 20[B] \text{ e } v_3 = 15[A] + 200[B] \tag{2.170}$$

as tensões v_1, v_2 e v_3 são dadas em mV e as concentrações em $milimol/l$.

a) Encontre uma expressão para os valores de $[A]$ e $[B]$ em termos de v_1 e v_2 e, depois, em termos de v_1 e v_3. Dica: Se $ad \neq bc$, então a matriz A dada por

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{2.171}$$

é inversível e sua inversa é dada por

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{2.172}$$

b) Sabendo que incerteza relativa associada às sensibilidades dos sensores 1 e 2 é de 2% e que a incerteza relativa associada às sensibilidades do sensor 3 é 10%, verifique a incerteza associada à medida feita com o par $1-2$ e o par $1-3$. Use $[A] = [B] = 10milimol/l$. Dica: Você deve diferenciar as grandezas $[A]$ e $[B]$ em relação aos valores das tensões.

Solução. Em ambos casos, temos a seguinte estrutura:

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} [A] \\ [B] \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{2.173}$$

De forma que

$$\begin{bmatrix} [A] \\ [B] \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{1}{S_{11}S_{22} - S_{12}S_{21}} \begin{bmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{2.174}$$

Portanto

$$[A] = \frac{S_{22}v_1 - S_{12}v_2}{S_{11}S_{22} - S_{12}S_{21}} \tag{2.175}$$

$$[B] = \frac{-S_{21}v_1 + S_{11}v_2}{S_{11}S_{22} - S_{12}S_{21}} \tag{2.176}$$

Usando derivação logarítmica, temos

$$\frac{1}{[A]}\frac{\partial[A]}{\partial S_{11}} = -\frac{S_{22}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.177}$$

$$\frac{1}{[A]}\frac{\partial[A]}{\partial S_{12}} = -\frac{v_2}{S_{22}v_1-S_{12}v_2}+\frac{S_{21}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.178}$$

$$= -\frac{[A]}{[B]}\cdot\frac{S_{22}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.179}$$

$$\frac{1}{[A]}\frac{\partial[A]}{\partial S_{21}} = \frac{S_{12}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.180}$$

$$\frac{1}{[A]}\frac{\partial[A]}{\partial S_{22}} = \frac{v_1}{S_{22}v_1-S_{12}v_2}-\frac{S_{11}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.181}$$

$$= \frac{[A]}{[B]}\cdot\frac{S_{12}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.182}$$

e

$$\frac{1}{[B]}\frac{\partial[B]}{\partial S_{11}} = \frac{v_2}{-S_{21}v_1+S_{11}v_2}-\frac{S_{22}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.183}$$

$$= \frac{[B]}{[A]}\frac{S_{21}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.184}$$

$$\frac{1}{[B]}\frac{\partial[B]}{\partial S_{12}} = \frac{S_{21}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.185}$$

$$\frac{1}{[B]}\frac{\partial[B]}{\partial S_{21}} = -\frac{v_1}{-S_{21}v_1+S_{11}v_2}+\frac{S_{21}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.186}$$

$$= -\frac{[B]}{[A]}\frac{S_{11}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.187}$$

$$\frac{1}{[B]}\frac{\partial[B]}{\partial S_{22}} = -\frac{S_{11}}{S_{11}S_{22}-S_{12}S_{21}} \tag{2.188}$$

$$\tag{2.189}$$

E o erro associado às medidas pode ser aproximado por

$$\frac{1}{[A]}\delta_{[A]} = \left|\frac{1}{[A]}\frac{\partial[A]}{\partial S_{11}}\right|\delta_{S_{11}}+\left|\frac{1}{[A]}\frac{\partial[A]}{\partial S_{12}}\right|\delta_{S_{12}} \tag{2.190}$$

$$+ \left|\frac{1}{[A]}\frac{\partial[A]}{\partial S_{21}}\right|\delta_{S_{21}}+\left|\frac{1}{[A]}\frac{\partial[A]}{\partial S_{22}}\right|\delta_{S_{22}} \tag{2.191}$$

$$= \frac{1}{|\det S|}\left[S_{22}\delta_{S_{11}}+\frac{[A]}{[B]}S_{22}\delta_{S_{12}}+S_{12}\delta_{S_{21}}+\frac{[A]}{[B]}S_{12}\delta_{S_{22}}\right] \tag{2.192}$$

Analogamente, temos:

$$\frac{1}{[B]}\delta_{[B]} = \frac{1}{|\det S|}\left[\frac{[B]}{[A]}S_{21}\delta_{S_{11}} + S_{21}\delta_{S_{11}} + \frac{[B]}{[A]}S_{11}\delta_{S_{21}} + S_{11}\delta_{S_{22}}\right] \quad (2.193)$$

onde não se indicou $|S_{ij}|$ nem $|[.]|$ pois são todos positivos.

Fazemos agora a aplicação numérica:

Caso do par 1-2:

$$\det S = \begin{vmatrix} 270 & 30 \\ 140 & 20 \end{vmatrix} = 1200 \quad (2.194)$$

$$\begin{aligned}
\frac{1}{[A]}\delta_{[A]} &= \frac{1}{1200}\left[20 \times 270 \times 2\% + 20 \times 30 \times 2\%\right. && (2.195)\\
&\quad \left. + 30 \times 140 \times 2\% + 30 \times 20 \times 2\%\right] && (2.196)\\
&= \frac{216}{1200} = 0.18 = 18\% && (2.197)\\
\frac{1}{[B]}\delta_{[B]} &= \frac{1}{1200}\left[140 \times 270 \times 2\% + 140 \times 30 \times 2\%\right. && (2.198)\\
&\quad \left. + 270 \times 140 \times 2\% + 270 \times 20 \times 2\%\right] && (2.199)\\
&= \frac{426}{1200} = 0.355 = 35.5\% && (2.200)
\end{aligned}$$

Caso do par 1-3:

$$\det S = \begin{vmatrix} 270 & 30 \\ 15 & 200 \end{vmatrix} = 53550 \quad (2.201)$$

$$\begin{aligned}
\frac{1}{[A]}\delta_{[A]} &= \frac{1}{53550}\left[200 \times 270 \times 2\% + 200 \times 30 \times 2\%\right. && (2.202)\\
&\quad \left. + 30 \times 15 \times 10\% + 30 \times 200 \times 10\%\right] && (2.203)\\
&= \frac{1804{,}6}{52550} \approx 0.0337 = 3.37\% && (2.204)\\
\frac{1}{[B]}\delta_{[B]} &= \frac{1}{53550}\left[15 \times 270 \times 2\% + 15 \times 30 \times 2\%\right. && (2.205)\\
&\quad \left. + 270 \times 15 \times 10\% + 270 \times 200 \times 10\%\right] && (2.206)\\
&= \frac{5895}{53550} \approx 0.11 = 11\% && (2.207)
\end{aligned}$$

Conclusão, apesar de o sensor 3 apresentar uma incerteza cinco vezes maior na sensibilidade, a escolha do sensor 3 para fazer par ao sensor 1 parece mais adequada.

$\Diamond$

Exercícios

E 2.9.1. Considere as expressões:

$$\frac{\exp(1/\mu)}{1+\exp(1/\mu)} \tag{2.208}$$

e

$$\frac{1}{\exp(-1/\mu)+1} \tag{2.209}$$

com $\mu > 0$. Verifique que elas são idênticas como funções reais. Teste no computador cada uma delas para $\mu = 0{,}1$, $\mu = 0{,}01$ e $\mu = 0{,}001$. Qual dessas expressões é mais adequada quando μ é um número pequeno? Por quê?

E 2.9.2. Encontre expressões alternativas para calcular o valor das seguintes funções quando x é próximo de zero.

a) $f(x) = \frac{1-\cos(x)}{x^2}$

b) $g(x) = \sqrt{1+x} - 1$

c) $h(x) = \sqrt{x+10^6} - 10^3$

d) $i(x) = \sqrt{1+e^x} - \sqrt{2}$ Dica: Faça $y = e^x - 1$

E 2.9.3. Use uma identidade trigonométrica adequada para mostrar que:

$$\frac{1-\cos(x)}{x^2} = \frac{1}{2}\left(\frac{\operatorname{sen}(x/2)}{x/2}\right)^2. \tag{2.210}$$

Analise o desempenho destas duas expressões no computador quando x vale 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-200} e 0. Discuta o resultado. **Dica:** Para $|x| < 10^{-5}$, $f(x)$ pode ser aproximada por $1/2 - x^2/24$ com erro de truncamento inferior a 10^{-22}.

E 2.9.4. Reescreva as expressões:

$$\sqrt{e^{2x}+1} - e^x \qquad \text{e} \qquad \sqrt{e^{2x}+x^2} - e^x \tag{2.211}$$

de modo que seja possível calcular seus valores para $x = 100$ utilizando a aritmética de ponto flutuante ("Double") no computador.

Prof. M.e Daniel Cassimiro

E 2.9.5. Na teoria da relatividade restrita, a energia cinética de uma partícula e sua velocidade se relacionam pela seguinte fórmula:

$$E = mc^2 \left(\frac{1}{\sqrt{1-(v/c)^2}} - 1 \right), \tag{2.215}$$

onde E é a energia cinética da partícula, m é a massa de repouso, v o módulo da velocidade e c a velocidade da luz no vácuo dada por $c = 299792458m/s$. Considere que a massa de repouso $m = 9{,}10938291 \times 10^{-31}kg$ do elétron seja conhecida com erro relativo de 10^{-9}. Qual é o valor da energia e o erro relativo associado a essa grandeza quando $v = 0{,}1c$, $v = 0{,}5c$, $v = 0{,}99c$ e $v = 0{,}999c$ sendo que a incerteza relativa na medida da velocidade é 10^{-5}?

Chapter 3

Solução de equações de uma variável

Neste capítulo, construiremos aproximações numéricas para a solução de **equações algébricas em uma única variável real**. Observamos que obter uma solução para uma dada equação é equivalente a encontrar um **zero de uma função real** apropriada. Com isso, iniciamos este capítulo discutindo condições de existência e unicidade de raízes de funções de uma variável real. Então, apresentamos o **método da bisseção** como uma primeira abordagem numérica para a solução de tais equações.

Em seguida, exploramos outra abordagem via **iteração do ponto fixo**. Desta, obtemos o **método de Newton**[1], para o qual estudamos aplicações e critérios de convergência. Por fim, apresentamos o **método das secantes** como uma das possíveis variações do método de Newton.

3.1 Existência e unicidade

O **teorema de Bolzano**[2] nos fornece condições suficientes para a existência do zero de uma função. Este é uma aplicação direta do **teorema do valor intermediário**.

Teorema 3.1.1 (Teorema de Bolzano)**.** *Se $f : [a, b] \to \mathbb{R}$, $y = f(x)$, é uma função contínua tal que $f(a) \cdot f(b) < 0$*[3]*, então existe $x^* \in (a, b)$ tal que $f(x^*) = 0$.*

Proof. O resultado é uma consequência imediata do teorema do valor intermediário que estabelece que dada uma função contínua $f : [a, b] \to \mathbb{R}$, $y = f(x)$, tal que

[1] Sir Isaac Newton, 1642 - 1727, matemático e físico inglês.

[2] Bernhard Placidus Johann Gonzal Nepomuk Bolzano, 1781 - 1848, matemático do Reino da Boêmia.

[3] Esta condição é equivalente a dizer que a função troca de sinal no intervalo.

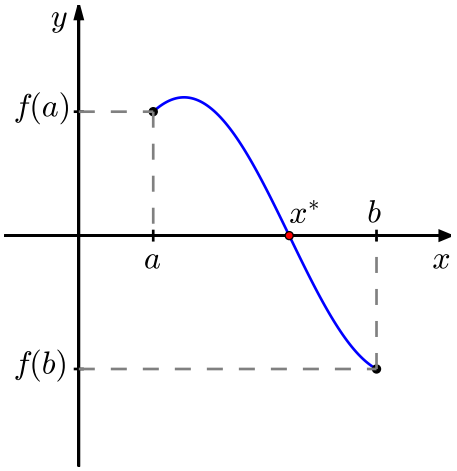


Figure 3.1: Teorema de Bolzano.

$f(a) < f(b)$ (ou $f(b) < f(a)$), então para qualquer $d \in (f(a), f(b))$ (ou $k \in (f(b), f(a))$) existe $x^* \in (a, b)$ tal que $f(x^*) = k$. Ou seja, nestas notações, se $f(a) \cdot f(b) < 0$, então $f(a) < 0 < f(b)$ (ou $f(b) < 0 < f(a)$). Logo, tomando $k = 0$, temos que existe $x^* \in (a, b)$ tal que $f(x^*) = k = 0$. □

Em outras palavras, se $f(x)$ é uma função contínua em um dado intervalo no qual ela troca de sinal, então ela tem pelo menos um zero neste intervalo (veja a Figura 3.1).

Exemplo 3.1.1. Mostre que existe pelo menos uma solução da equação $e^x = x+2$ no intervalo $(-2,0)$.

Solução. Primeiramente, observamos que resolver a equação $e^x = x+2$ é equivalente a resolver $f(x) = 0$ com $f(x) = e^x - x - 2$. Agora, como $f(-2) = e^{-2} > 0$ e $f(0) = -1 < 0$, temos do teorema de Bolzano que existe pelo menos um zero de $f(x)$ no intervalo $(-2, 0)$. E, portanto, existe pelo menos uma solução da equação dada no intervalo $(-2, 0)$.

Podemos usar o `GNU Octave` para estudar esta função. Por exemplo, podemos definir a função $f(x)$ e computá-la nos extremos do intervalo dado com os seguintes comandos:

```
>> f = @(x) exp(x)-x-2
f = f(x) = exp(x)-x-2
>> f(-2),f(0)
ans =  0.13534
ans = -1
```

Alternativamente (e com maior precisão), podemos verificar diretamente o sinal da função nos pontos desejados com comando `sign`:

```
>> sign(f(-2)),sign(f(0))
ans =  1
ans = -1
```

$\Diamond$

Quando procuramos aproximações para zeros de funções, é aconselhável isolar cada raiz em um intervalo. Desta forma, gostaríamos de poder garantir a existência e a unicidade da raiz dentro de um dado intervalo. A seguinte proposição nos fornece condições suficientes para tanto.

Proposição 3.1.1. *Se $f : [a,b] \to \mathbb{R}$ é um função diferenciável, $f(a) \cdot f(b) < 0$ e $f'(x) > 0$ (ou $f'(x) < 0$) para todo $x \in (a,b)$, então existe um único $x^* \in (a,b)$ tal que $f(x^*) = 0$.*

Em outras palavras, para garantirmos que exista um único zero de uma dada função diferenciável em um intervalo, é suficiente que ela troque de sinal e seja monótona neste intervalo.

Exemplo 3.1.2. No Exemplo 3.1.1, mostramos que existe pelo menos um zero de $f(x) = e^x - x - 2$ no intervalo $(-2{,}0)$, pois $f(x)$ é contínua e $f(-2) \cdot f(0) < 0$. Agora, observamos que, além disso, $f'(x) = e^x - 1$ e, portanto, $f'(x) < 0$ para todo $x \in (-2{,}0)$. Logo, da Proposição 3.1.1, temos garantida a existência de um único zero no intervalo dado.

Podemos inspecionar o comportamento da função $f(x) = e^x - x - 2$ e de sua derivada fazendo seus gráficos no `GNU Octave`. Para tanto, podemos fazer o seguinte teste:

```
>> xx = linspace(-2,0,50);
>> f = @(x) exp(x)-x-2  #define f
f = f(x) = exp(x)-x-2
>> plot(xx,f(xx))grid on;hold on #gráfico de f
>> fl = @(x) exp(x)-1  #a derivada
fl = f(x) = exp(x)-1
>> plot(xx,fl(xx)) #gráfico de f'
```

A discussão feita nesta seção, especialmente o teorema de Bolzano, nos fornece os fundamentos para o método da bisseção, o qual discutimos na próxima seção.

Exercícios

E 3.1.1. Mostre que $\cos x = x$ tem solução no intervalo $[0, \pi/2]$.

E 3.1.2. Mostre que $\cos x = x$ tem uma única solução no intervalo $[0, \pi/2]$.

E 3.1.3. Interprete a equação $\cos(x) = kx$ como o problema de encontrar a intersecção da curva $y = \cos(x)$ com $y = kx$. Encontre o valor positivo k para o qual essa equação admite exatamente duas raízes positivas distintas.

E 3.1.4. Mostre que a equação:

$$\ln(x) + x^3 - \frac{1}{x} = 10 \tag{3.1}$$

possui uma única solução positiva.

E 3.1.5. Use o teorema de Bolzano para mostrar que o erro absoluto ao aproximar o zero da função $f(x) = e^x - x - 2$ por $\overline{x} = -1{,}841$ é menor que 10^{-3}.

E 3.1.6. Mostre que o erro absoluto associado à aproximação $\overline{x} = 1{,}962$ para a solução exata x^* de:

$$e^x + \operatorname{sen}(x) + x = 10 \tag{3.2}$$

é menor que 10^{-4}.

E 3.1.7. Mostre que a equação

$$\ln(x) + x - \frac{1}{x} = v \tag{3.3}$$

possui uma solução para cada v real e que esta solução é única.

3.2 Método da bisseção

O **método da bisseção** explora o fato de que uma função contínua $f : [a, b] \to \mathbb{R}$ com $f(a) \cdot f(b) < 0$ tem um zero no intervalo (a, b) (veja o teorema de Bolzano 3.1.1). Assim, a ideia para aproximar o zero de uma tal função $f(x)$ é tomar, como aproximação inicial, o ponto médio do intervalo $[a, b]$, isto é:

$$x^{(0)} = \frac{(a+b)}{2}. \tag{3.4}$$

Pode ocorrer de $f(x^{(0)}) = 0$ e, neste caso, o zero de $f(x)$ é $x^* = x^{(0)}$. Caso contrário, se $f(a) \cdot f(x^{(0)}) < 0$, então $x^* \in (a, x^{(0)})$. Neste caso, tomamos como

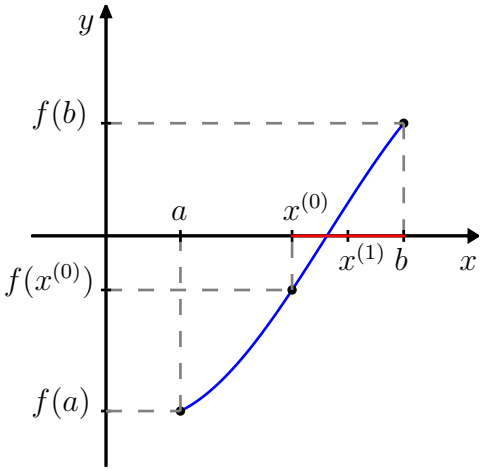


Figure 3.2: Método da bisseção.

nova aproximação do zero de $f(x)$ o ponto médio do intervalo $[a, x^{(0)}]$, isto é, $x^{(1)} = (a + x^{(0)})/2$. No outro caso, temos $f(x^{(0)}) \cdot f(b) < 0$ e, então, tomamos $x^{(1)} = (x^{(0)} + b)/2$. Repetimos este procedimento até obtermos a aproximação desejada (veja Figura 3.2).

De forma mais precisa, suponha que queiramos calcular uma aproximação com uma certa precisão TOL para um zero x^* de uma dada função contínua $f : [a, b] \to \mathbb{R}$ tal que $f(a) \cdot f(b) < 0$. Iniciamos, tomando $n = 0$ e:

$$a^{(n)} = a, \quad b^{(n)} = b \quad \text{e} \quad x^{(n)} = \frac{a^{(n)} + b^{(n)}}{2}. \tag{3.5}$$

Verificamos o **critério de parada**, isto é, se $f(x^{(n)}) = 0$ ou:

$$\frac{|b^{(n)} - a^{(n)}|}{2} < TOL, \tag{3.6}$$

então $x^{(n)}$ é a aproximação desejada. Caso contrário, preparamos a próxima iteração $n + 1$ da seguinte forma: se $f(a^{(n)}) \cdot f(x^{(n)}) < 0$, então definimos $a^{(n+1)} = a^{(n)}$ e $b^{(n+1)} = x^{(n)}$; no outro caso, se $f(x^{(n)}) \cdot f(b^{(n)}) < 0$, então definimos $a^{(n+1)} = x^{(n)}$ e $b^{(n+1)} = b^{(n)}$. Trocando n por $n + 1$, temos a nova aproximação do zero de $f(x)$ dada por:

$$x^{(n+1)} = \frac{a^{(n+1)} + b^{(n+1)}}{2}. \tag{3.7}$$

Voltamos a verificar o critério de parada acima e, caso não satisfeito, iteramos novamente. Iteramos até obtermos a aproximação desejada ou o número máximo de iterações ter sido atingido.

Table 3.1: Iteração do método da bisseção para o Exemplo 3.2.1.

n	$a^{(n)}$	$b^{(n)}$	$x^{(n)}$	$f(a^{(n)})f(x^{(n)})$	$\frac{\lvert b^{(n)} - a^{(n)}\rvert}{2}$
0	-2	0	-1	< 0	1
1	-2	-1	$-1{,}5$	< 0	0,5
2	-2	$-1{,}5$	$-1{,}75$	< 0	0,25
3	-2	$-1{,}75$	$-1{,}875$	> 0	0,125
4	$-1{,}875$	$-1{,}75$	$-1{,}8125$	< 0	0,0625

Exemplo 3.2.1. Use o método da bisseção para calcular uma solução de $e^x = x+2$ no intervalo $[-2, 0]$ com precisão $TOL = 10^{-1}$.

Solução. Primeiramente, observamos que resolver a equação dada é equivalente a calcular o zero de $f(x) = e^x - x - 2$. Além disso, temos $f(-2) \cdot f(0) < 0$. Desta forma, podemos iniciar o método da bisseção tomando o intervalo inicial $[a^{(0)}, b^{(0)}] = [-2, 0]$ e:

$$x^{(0)} = \frac{a^{(0)} + b^{(0)}}{2} = -1. \tag{3.8}$$

Apresentamos as iterações na Tabela 3.1. Observamos que a precisão $TOL = 10^{-1}$ foi obtida aproximando o zero de $f(x)$ por $x^{(4)} = -1{,}8125$.

Usando o `GNU Octave` neste exemplo, temos:

```
>> f = @(x) exp(x) - x - 2
f = f(x) = exp(x) - x - 2
>> a=-2, b=0, x=(a+b)/2, TOL = (b-a)/2, sign(f(a)*f(x))
a = -2
b = 0
x = -1
TOL =  1
ans = -1
>> b=x, x=(a+b)/2, TOL = (b-a)/2, sign(f(a)*f(x))
b = -1
x = -1.5000
TOL =  0.50000
ans = -1
```

e, assim, sucessivamente. Veja o código completo na Seção 3.2.1. ◊

Vamos agora discutir sobre a **convergência** do método da bisseção, que é garantida pelo Teorema 3.2.1.

Teorema 3.2.1 (Convergência do método da bisseção)**.** *Sejam $f : [a, b] \to \mathbb{R}$ uma função contínua tal que $f(a) \cdot f(b) < 0$ e x^* o único zero de $f(x)$ no intervalo (a, b). Então, a sequência $\{x^{(n)}\}_{n\geq 0}$ do método da bisseção satisfaz:*

$$|x^{(n)} - x^*| < \frac{b-a}{2^{n+1}}, \quad \forall n \geq 0. \tag{3.9}$$

Em particular, $x^{(n)} \to x^*$ quando $n \to \infty$.

Proof. Notemos que, a cada iteração, a distância entre a aproximação $x^{(n)}$ e o zero x^* da função é menor ou igual que a metade do tamanho do intervalo $[a^{(n)}, b^{(n)}]$ (veja Figura 3.2), isto é:

$$|x^{(n)} - x^*| \leq \frac{b^{(n)} - a^{(n)}}{2}. \tag{3.10}$$

Por construção do método, temos $[a^{(n)}, b^{(n)}] \subset [a^{(n-1)}, b^{(n-1)}]$ quando $n \geq 1$ e:

$$b^{(n)} - a^{(n)} = \frac{b^{(n-1)} - a^{(n-1)}}{2}. \tag{3.11}$$

Desta forma:

$$|x^{(n)} - x^*| \leq \frac{b^{(n)} - a^{(n)}}{2} = \frac{b^{(n-1)} - a^{(n-1)}}{2^2} = \cdots = \frac{b^{(0)} - a^{(0)}}{2^{n+1}}, \quad \forall n \geq 1. \tag{3.12}$$

Logo, vemos que:

$$|x^{(n)} - x^*| \leq \frac{b-a}{2^{n+1}}, \quad \forall n \geq 0. \tag{3.13}$$

□

Observamos que a hipótese de que $f(x)$ tenha um único zero no intervalo não é realmente necessária. Se a função tiver mais de um zero no intervalo inicial, as iterações ainda convergem para um dos zeros. Veja o Exercício 3.2.3.

Observação 3.2.1. O Teorema 3.2.1 nos fornece uma estimativa para a convergência do método da bisseção. Aproximadamente, temos:

$$|x^{(n+1)} - x^*| \lesssim \frac{1}{2}|x^{(n)} - x^*|. \tag{3.14}$$

Isto nos leva a concluir que o método da bisseção tem **taxa de convergência** linear.

Exemplo 3.2.2. No Exemplo 3.2.1, depois de escolhida a aproximação inicial $x^{(0)} = -1 \in [a, b] = [-2, 0]$, precisamos de 4 iterações do método da bisseção para computar uma aproximação com precisão de 10^{-1} do zero de $f(x) = e^x - x - 2$. Poderíamos ter estimado o número de iterações **a priori**, pois, como vimos acima:

$$|x^{(n)} - x^*| \leq \frac{b-a}{2^{n+1}}, \quad n \geq 0. \tag{3.15}$$

Logo, temos:

$$\begin{aligned} |x^{(n)} - x^*| &< \frac{b-a}{2^{n+1}} = \frac{2}{2^{n+1}} && (3.16) \\ &= 2^{-n} < 10^{-1} \Rightarrow n > -\log_2 10^{-1} \approx 3{,}32. && (3.17) \end{aligned}$$

Isto está de acordo com o experimento numérico realizado naquele exemplo.

O método da bisseção tem a boa propriedade de garantia de convergência, bem como de fornecer uma simples estimativa do erro na aproximação calculada. Entretanto, a taxa de convergência linear é superada por outros métodos. A construção de tais métodos está, normalmente, associada à iteração do ponto fixo, a qual exploramos na próxima seção.

3.2.1 Código GNU Octave: método da bisseção

O seguinte código é uma implementação no `GNU Octave` do algoritmo da bisseção. As variáveis de entrada são:

- `f` - função objetivo
- `a` - extremo esquerdo do intervalo de inspeção $[a, b]$
- `b` - extremo direito do intervalo de inspeção $[a, b]$
- `TOL` - tolerância para o erro absoluto (critério de parada)
- `N` - número máximo de iterações depois da aproximação inicial

A variável de saída é:

- `p` - aproximação de uma raiz x^* de `f`, satisfazendo $|p - x^*| < TOL$ ou $f(p) = 0$.

```
function [p] = bissecao(f, a, b, TOL, N)
  i = 1;
  fa = f(a);
  while (i <= N)
```

```
    #iteracao da bissecao
    p = a + (b-a)/2;
    fp = f(p);
    #condicao de parada
    if ((fp == 0) || ((b-a)/2 < TOL))
      return;
    endif
    #bissecta o intervalo
    i = i+1;
    if (fa * fp > 0)
      a = p;
      fa = fp;
    else
      b = p;
    endif
  endwhile
  error ('Num. max. de iter. excedido!');
endfunction
```

Exercícios

E 3.2.1. Considere a equação $\sqrt{x} = \cos(x)$. Use o método da bisseção com intervalo inicial $[a, b] = [0, 1]$ e $x^{(1)} = (a + b)/2$ para calcular a aproximação $x^{(4)}$ da solução desta equação.

E 3.2.2. Trace o gráfico e isole as três primeiras raízes positivas da função:

$$f(x) = 5\,\mathrm{sen}\,(x^2) - \exp\left(\frac{x}{10}\right) \tag{3.18}$$

em intervalos de comprimento 0,1. Então, use o método da bisseção para obter aproximações dos zeros desta função com precisão de 10^{-5}.

E 3.2.3. O polinômio $p(x) = -4+8x-5x^2+x^3$ tem raízes $x_1 = 1$ e $x_2 = x_3 = 2$ no intervalo $[1/2, 3]$.

a) Se o método da bisseção for usando com o intervalo inicial $[1/2, 3]$, para qual raiz as iterações convergem?

b) É possível usar o método da bisseção para a raiz $x = 2$? Justifique sua resposta.

E 3.2.4. O polinômio $f(x) = x^4 - 4x^2 + 4$ possui raízes duplas em $\sqrt{2}$ e $-\sqrt{2}$. O método da bisseção pode ser aplicado a f? Explique.

E 3.2.5. Mostre que a equação do Problema 3.1.7 possui uma solução no intervalo $[1, v+1]$ para todo v positivo. Dica: defina $f(x) = \ln(x) + x - \frac{1}{x} - v$ e considere a seguinte estimativa:

$$f(v+1) = f(1) + \int_1^{v+1} f'(x)dx \geq -v + \int_1^{v+1} dx = 0. \tag{3.19}$$

Use esta estimativa para iniciar o método de bisseção e obtenha o valor da raiz com pelo menos 6 algarismos significativos para $v = 1, 2, 3, 4$ e 5.

E 3.2.6. (Estática) Considere o seguinte problema físico: uma plataforma está fixa a uma parede através de uma dobradiça cujo momento é dado por:

$$\tau = k\theta, \tag{3.20}$$

onde θ é ângulo da plataforma com a horizontal e k é uma constante positiva. A plataforma é feita de material homogêneo, seu peso é P e sua largura é l. Modele a relação entre o ângulo θ e o peso P próprio da plataforma. Encontre o valor de θ quando $l = 1$ m, $P = 200$ N, $k = 50$ Nm/rad, sabendo que o sistema está em equilíbrio. Use o método da bisseção e expresse o resultado com 4 algarismos significativos.

E 3.2.7. Considere a equação de Lambert dada por:

$$xe^x = t, \tag{3.21}$$

onde t é um número real positivo. Mostre que esta equação possui uma única solução x^* que pertence ao intervalo $[0, t]$. Usando esta estimativa como intervalo inicial, quantos passos são necessário para obter o valor numérico de x^* com erro absoluto inferior a 10^{-6} quando $t = 1$, $t = 10$ e $t = 100$ através do método da bisseção? Obtenha esses valores.

E 3.2.8. (Eletrônica) O desenho abaixo mostra um circuito não linear envolvendo uma fonte de tensão constante, um diodo retificador e um resistor. Sabendo que a relação entre a corrente (I_d) e a tensão (v_d) no diodo é dada pela seguinte expressão:

$$I_d = I_R \left(\exp\left(\frac{v_d}{v_t}\right) - 1 \right), \tag{3.22}$$

onde I_R é a corrente de condução reversa e v_t, a tensão térmica dada por $v_t = \frac{kT}{q}$ com k, a constante de Boltzmann, T a temperatura de operação e q, a carga do

elétron. Aqui $I_R = 1pA = 10^{-12}$ A, $T = 300$ K. Escreva o problema como uma equação na incógnita v_d e, usando o método da bisseção, resolva este problema com 3 algarismos significativos para os seguintes casos:

a) $V = 30$ V e $R = 1$ kΩ.

b) $V = 3$ V e $R = 1$ kΩ.

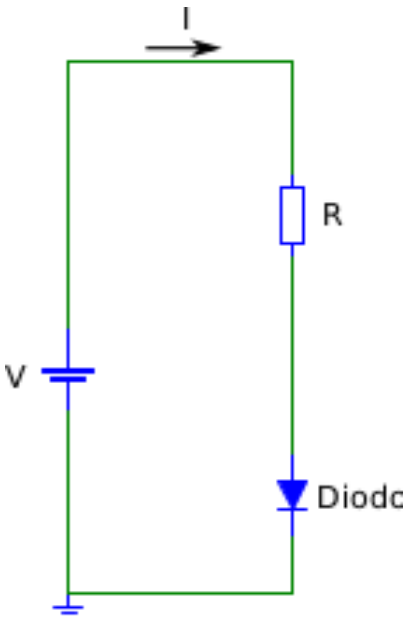
c) $V = 3$ V e $R = 10$ kΩ.

d) $V = 300$ mV e $R = 1$ kΩ.

e) $V = -300$ mV e $R = 1$ kΩ.

f) $V = -30$ V e $R = 1$ kΩ.

g) $V = -30$ V e $R = 10$ kΩ.



Dica: $V = RI_d + v_d$.

E 3.2.9. (Propagação de erros) Obtenha os valores de I_d no Problema 3.2.8. Lembre que existem duas expressões disponíveis:

$$I_d = I_R \left(\exp\left(\frac{v_d}{v_t}\right) - 1 \right) \tag{3.23}$$

e

$$I_d = \frac{v - v_d}{R} \tag{3.24}$$

Faça o estudo da propagação do erro e decida qual a melhor expressão em cada caso.

3.3 Iteração de ponto fixo

Nesta seção, discutimos a abordagem da **iteração do ponto fixo** para a solução numérica de equações de uma variável real. Observamos que sempre podemos reescrever uma equação da forma $f(x) = 0$ (problema de encontrar os zeros de uma função) em uma equação equivalente na forma $g(x) = x$ (**problema de ponto fixo**). Um ponto $x = x^*$ tal que $g(x^*) = x^*$ é chamado de **ponto fixo** da função $g(x)$. Geometricamente, um ponto fixo de uma função é um ponto de interseção entre a reta $y = x$ com o gráfico da função $g(x)$ (veja Figura 3.3).

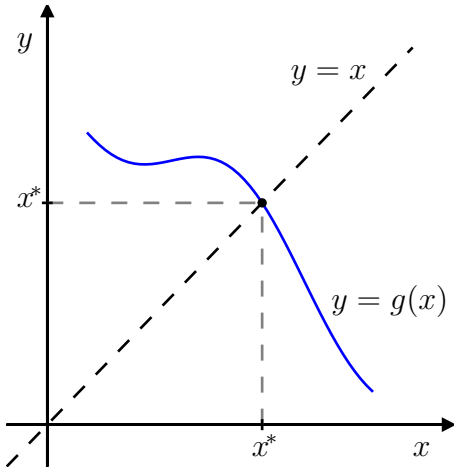


Figure 3.3: Ponto fixo $g(x^*) = x^*$.

Exemplo 3.3.1. Resolver a equação $e^x = x + 2$ é equivalente a resolver $f(x) = 0$, com $f(x) = e^x - x - 2$. Estes são equivalentes a resolver $g(x) = x$, com $g(x) = e^x - 2$, isto é:

$$e^x = x + 2 \Leftrightarrow e^x - x - 2 = 0 \Leftrightarrow e^x - 2 = x \tag{3.25}$$

Dada uma função $g(x)$, a **iteração do ponto fixo** consiste em computar a seguinte sequência recursiva:

$$x^{(n+1)} = g(x^{(n)}), \quad n \geq 1, \tag{3.26}$$

onde $x^{(1)}$ é uma aproximação inicial do ponto fixo.

Exemplo 3.3.2 (Método babilônico)**.** O método babilônico[4] é de uma iteração de ponto fixo para extrair a raiz quadrada de um número positivo A, isto é, resolver a equação $x^2 = A$.

Seja $r > 0$ uma aproximação para $\sqrt{A}$. Temos três possibilidades:

- $r > \sqrt{A} \Longrightarrow \frac{A}{r} < \sqrt{A} \Longrightarrow \sqrt{A} \in \left(\frac{A}{r}, r\right)$;
- $r = \sqrt{A} \Longrightarrow \frac{A}{r} = \sqrt{A}$;
- $r < \sqrt{A} \Longrightarrow \frac{A}{r} > \sqrt{A} \Longrightarrow \sqrt{A} \in \left(r, \frac{A}{r}\right)$.

[4]Heron de Alexandria, 10 d.C. - 70 d.C., matemático grego.

Ou seja, $\sqrt{A}$ sempre está no intervalo entre r e $\frac{A}{r}$, no qual podemos buscar uma nova aproximação como, por exemplo, pelo ponto médio:

$$x = \frac{r + \frac{A}{r}}{2}. \tag{3.27}$$

Aplicando esse método repetidas vezes, podemos construir a iteração (de ponto fixo):

$$x^{(1)} = r \tag{3.28}$$

$$x^{(n+1)} = \frac{x^{(n)}}{2} + \frac{A}{2x^{(n)}}, \quad n = 1,2,3,... \tag{3.29}$$

Por exemplo, para obter uma aproximação para $\sqrt{5}$, podemos iniciar com a aproximação inicial $r = 2$ e $A = 5$. Então, tomamos $x^{(1)} = 2$ e daí seguem as aproximações:

$$x^{(2)} = \frac{2}{2} + \frac{2,5}{2} = 2,25 \tag{3.30}$$

$$x^{(3)} = \frac{2,25}{2} + \frac{2,5}{2,25} = 2,2361111 \tag{3.31}$$

$$x^{(4)} = \frac{2,2361111}{2} + \frac{2,5}{2,2361111} = 2,236068 \tag{3.32}$$

$$x^{(5)} = \frac{2,236068}{2} + \frac{2,5}{2,236068} = 2,236068 \tag{3.33}$$

O método babilônico sugere que a iteração do ponto fixo pode ser uma abordagem eficiente para a solução de equações. Ficam, entretanto, as seguintes perguntas:

1. Será que a iteração do ponto fixo é convergente?

2. Caso seja convergente, será que o limite da sequência produzida, isto é, $x^* := \lim_{n\to\infty} x^{(n)}$ é um ponto fixo?

3. Caso seja convergente, qual é a taxa de convergência?

A segunda pergunta é a mais fácil de ser respondida. No caso de $g(x)$ ser contínua, se $x^{(n)} \to x^* \in (g)$, então:

$$x^* = \lim_{n\to\infty} x^{(n)} = \lim_{n\to\infty} g(x^{(n-1)}) = g\left(\lim_{n\to\infty} x^{(n-1)}\right) = g(x^*). \tag{3.34}$$

Antes de respondermos as outras perguntas acima, vejamos mais um exemplo.

Table 3.2: Iterações do ponto fixo para o Exemplo 3.3.3.

n	$x_1^{(n)}$	$x_2^{(n)}$
1	1,700	1,700
2	2,047	1,735
3	−0,8812	1,743
4	4,3013	1,746
5	−149,4	1,746

Exemplo 3.3.3. Considere o problema de encontrar o zero da função $f(x) = xe^x - 10$. Uma maneira geral de construir um problema de ponto fixo equivalente é o seguinte:

$$f(x) = 0 \Rightarrow \alpha f(x) = 0 \Rightarrow x - \alpha f(x) = x, \tag{3.35}$$

para qualquer parâmetro $\alpha \neq 0$. Consideremos, então, as seguintes duas funções:

$$g_1(x) = x - 0{,}5f(x) \quad \text{e} \quad g_2(x) = x - 0{,}05f(x). \tag{3.36}$$

Notamos que o ponto fixo destas duas funções coincide com o zero de $f(x)$. Construindo as iterações do ponto fixo:

$$x_1^{(n+1)} = g_1(x_1^{(n)}) \quad \text{e} \quad x_2^{(n+1)} = g_2(x_2^{(n)}), \tag{3.37}$$

tomando $x_1^{(1)} = x_2^{(1)} = 1{,}7$, obtemos os resultados apresentados na Tabela 3.2. Observamos que, enquanto, a iteração do ponto fixo com a função $g_1(x)$ ($\alpha = 0{,}5$) parece divergir, a iteração com a função $g_2(x)$ ($\alpha = 0{,}05$) parece convergir.

No `GNU Octave`, podemos computar as iterações do ponto fixo $x^{(n+1)} = g_1(x^{(n)})$ com o seguinte código:

```
>> f = @(x) x*exp(x)-10
f = f(x) = x*exp(x)-10
>> g1 = @(x) x - 0.5*f(x)
g1 = f(x) = x - 0.5*f(x)
>> x = 1.7;
>> x = g1(x)
x =  2.0471
>> x = g1(x)
x = -0.88119
```

e, assim, sucessivamente. Itere com a função $g_2(x)$ e verifique a convergência!

A fim de estudarmos a convergência da iteração do ponto fixo, apresentamos o teorema do ponto fixo.

3.3.1 Teorema do ponto fixo

O teorema do ponto fixo nos fornece condições suficientes para a existência e unicidade do ponto fixo, bem como para a convergência das iterações do método.

Definição 3.3.1. *Uma* ***contração*** *é uma função real* $g : [a, b] \to [a, b]$ *tal que:*

$$|g(x) - g(y)| \leq \beta|x - y|, \quad 0 \leq \beta < 1. \tag{3.38}$$

Observação 3.3.1. Seja $g : [a, b] \to [a, b]$, y=g(x).

- Se $g(x)$ é uma contração, então $g(x)$ é uma função contínua.
- Se $|g'(x)| < k$, $0 < k < 1$, para todo $x \in [a, b]$, então $g(x)$ é uma contração.

Teorema 3.3.1 (Teorema do ponto fixo)**.** *Se* $g : [a,b] \to [a,b]$ *é uma contração, então existe um único ponto* $x^* \in [a, b]$ *tal que* $g(x^*) = x^*$*, isto é,* x^* *é ponto fixo de* $g(x)$*. Além disso, a sequência* $\{x^{(n)}\}_{n\in\mathbb{N}}$ *dada por:*

$$x^{(n+1)} = g(x^{(n)}) \tag{3.39}$$

converge para x^* *para qualquer* $x^{(1)} \in [a, b]$*.*

Proof. Começamos demonstrando que existe pelo menos um ponto fixo. Para tal definimos a função $f(x) = x - g(x)$ e observamos que:

$$f(a) = a - g(a) \leq a - a = 0 \tag{3.40}$$

e

$$f(b) = b - g(b) \geq b - b = 0 \tag{3.41}$$

Se $f(a) = a$ ou $f(b) = b$, então o ponto fixo existe. Caso contrário, as desigualdades são estritas e a $f(x)$ muda de sinal no intervalo. Como esta função é contínua, pelo teorema de Bolzano 3.1.1, existe um ponto x^* no intervalo (a, b) tal que $f(x^*) = 0$, ou seja, $g(x^*) = x^*$. Isto mostra a existência.

Para provar que o ponto fixo é único, observamos que se x^* e x^{**} são pontos fixos, eles devem ser iguais, pois:

$$|x^* - x^{**}| = |g(x^*) - g(x^{**})| \leq \beta|x^* - x^{**}|. \tag{3.42}$$

A desigualdade $|x^* - x^{**}| \leq \beta|x^* - x^{**}|$ com $0 \leq \beta < 1$ implica $|x^* - x^{**}| = 0$.

Para demonstrar a convergência da sequência, observamos que:

$$|x^{(n+1)} - x^*| = |g(x^{(n)}) - x^*| = |g(x^{(n)}) - g(x^*)| \leq \beta|x^{(n)} - x^*|. \tag{3.43}$$

Daí, temos:

$$|x^{(n)} - x^*| \leq \beta|x^{(n-1)} - x^*| \leq \beta^2|x^{(n-2)} - x^*| \leq \cdots \leq \beta^n|x^{(0)} - x^*|. \tag{3.44}$$

Portanto, como $0 \leq \beta < 1$, temos:

$$\lim_{n\to\infty} |x^{(n)} - x^*| = 0, \tag{3.45}$$

ou seja, $x^{(n)} \to x^*$ quando $n \to \infty$. □

Observação 3.3.2. Do teorema do ponto fixo, temos que se $g(x)$ é uma contração com constante $0 \leq \beta < 1$, então:

$$|x^{(n+1)} - x^*| \leq \beta|x^{(n)} - x^*|, \quad n \geq 1. \tag{3.46}$$

Isto é, as iterações do ponto fixo têm taxa de convergência linear.

Exemplo 3.3.4. Mostre que o teorema do ponto fixo se aplica a função $g(x) = \cos(x)$ no intervalo $[1/2, 1]$, isto é, a iteração de ponto fixo converge para a solução da equação $\cos x = x$. Então, calcule as iterações do ponto fixo com aproximação inicial $x^{(1)} = 0{,}7$, estime o erro absoluto da aproximação e verifique a taxa de convergência.

Solução. Basta mostrarmos que:

a) $g([1/2{,}1]) \subseteq [1/2{,}1]$;

b) $|g'(x)| < \beta, \quad 0 < \beta < 1, \quad \forall x \in [1/2{,}1]$.

Para provar a), observamos que $g(x)$ é decrescente no intervalo, pelo que temos:

$$0{,}54 < \cos(1) \leq \cos(x) \leq \cos(1/2) < 0{,}88 \tag{3.47}$$

Como [0,54, 0,88] $\subseteq$ [0,5, 1], temos o item a).

Para provar o item b), observamos que:

$$g'(x) = -\operatorname{sen}(x). \tag{3.48}$$

Da mesma forma, temos a estimativa:

$$-0{,}85 < -\operatorname{sen}(1) \leq -\operatorname{sen}(x) \leq -\operatorname{sen}(1/2) < -0{,}47. \tag{3.49}$$

Assim, $|g'(x)| < 0{,}85$ e temos a desigualdade com $\beta = 0{,}85 < 1$.

A Tabela 3.3 apresenta o comportamento numérico da iteração do ponto fixo:

$$x^{(1)} = 0{,}7 \tag{3.50}$$

$$x^{(n+1)} = \cos(x^{(n)}), \quad n \geq 1. \tag{3.51}$$

n	$x^{(n)}$	$\epsilon_n := \lvert x^{(n)} - x^* \rvert$
1	0,70000	3,9E−02
2	0,76484	2,6E−02
3	0,72149	1,8E−02
4	0,75082	1,2E−02
5	0,73113	8,0E−03
6	0,74442	5,3E−03
7	0,73548	3,6E−03

Table 3.3: Iteração do ponto fixo para o Exemplo 3.3.4.

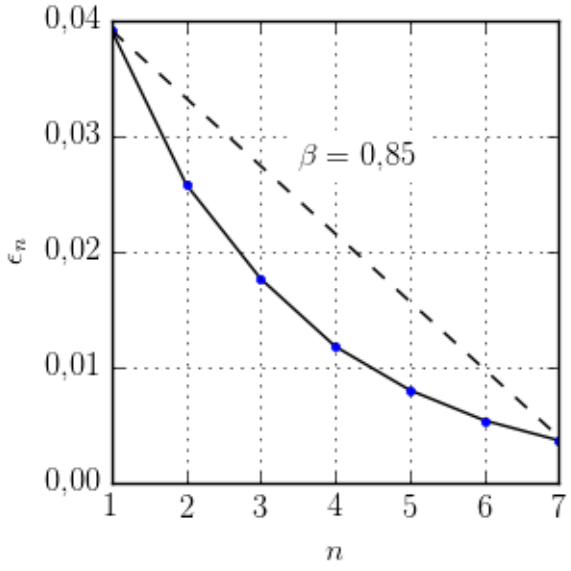


Figure 3.4: Decaimento do erro $\epsilon_n = |x^{(n)} - x^*|$ da iteração do ponto fixo estudada no Exemplo 3.3.4.

Para estimar o erro, consideramos $x^* = 0{,}7390851605$. A Figura 3.4 mostrar o decaimento do erro $\epsilon_n = |x^{(n)} - x^*|$ comparado com a taxa de convergência linear com $\beta = 0{,}85$.

No `GNU Octave`, podemos computar estas iterações e o erro absoluto com o seguinte código:

```
#est. da solucao
f = @(x) cos(x)-x;
xe = fsolve(f, 0.7);

#funcao do pto. fixo
g = @(x) cos(x);

#aprox. inicial
x0 = 0.7;
eps = abs(x0-xe);
printf("%1.5f %1.1e\n", x0, eps);

for i=2:7
  x = g(x0);
  eps = abs(x-xe);
  printf("%1.5f %1.1e\n", x, eps)
  x0 = x;
endfor
```

$\Diamond$

3.3.2 Teste de convergência

Seja $g : [a,b] \to \mathbb{R}$ uma função $C^0[a,b]$ e $x^* \in (a,b)$ um ponto fixo de g. Então x^* é dito estável se existe uma região $(x^* - \delta, x^* + \delta)$ chamada bacia de atração tal que $x^{(n+1)} = g(x^{(n)})$ é convergente sempre que $x^{(0)} \in (x^* - \delta, x^* + \delta)$.

Proposição 3.3.1 (Teste de convergência)**.** *Se $g \in C^1[a,b]$ e $|g'(x^*)| < 1$, então x^* é estável. Se $|g'(x^*)| > 1$ é instável e o teste é inconclusivo quando $|g'(x^*)| = 1$.*

Exemplo 3.3.5. No Exemplo 3.3.3, observamos que a função $g_1(x)$ nos forneceu uma iteração divergente, enquanto que a função $g_2(x)$ forneceu uma iteração convergente (veja a Figura 3.5. Estes comportamentos são explicados pelo teste da convergência. Com efeito, sabemos que o ponto fixo destas funções está no intervalo $[1{,}6, 1{,}8]$ e temos:

$$|g_1'(x)| = |1 - 0{,}5(x+1)e^x| > 4{,}8, \quad \forall x \in [1{,}6, 1{,}8], \tag{3.52}$$

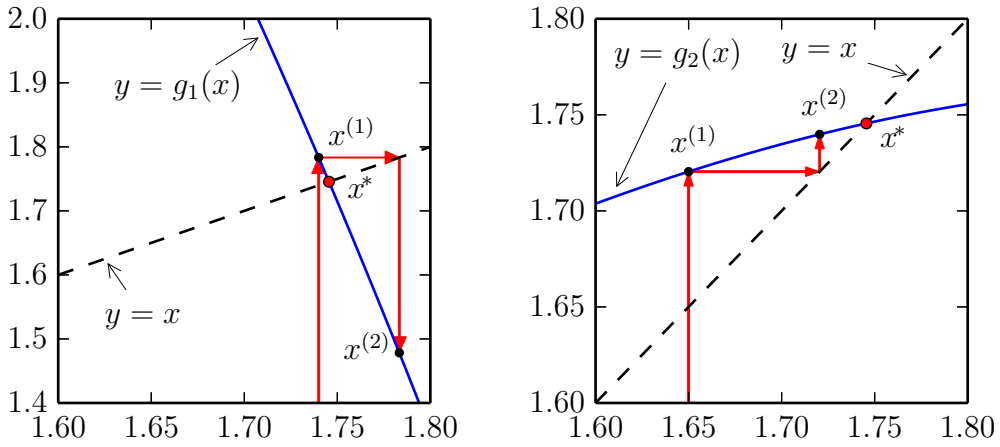


Figure 3.5: Ilustração das iterações do ponto fixo para: (esquerda) $y = g_1(x)$ e (direita) $y = g_2(x)$. Veja Exemplo 3.3.5.

enquanto:

$$|g_2'(x)| = |1 - 0{,}05(x+1)e^x| < 0{,}962, \quad \forall x \in [1{,}6, 1{,}8]. \tag{3.53}$$

3.3.3 Estabilidade e convergência

A fim de compreendermos melhor os conceitos de estabilidade e convergência, considere uma função $\Phi(x)$ com um ponto fixo $x^* = g(x^*)$ e analisemos o seguinte processo iterativo:

$$x^{(n+1)} = g\left(x^{(n)}\right) \tag{3.54}$$

$$x^{(0)} = x \tag{3.55}$$

Vamos supor que a função $g(x)$ pode ser aproximada por seu polinômio de Taylor em torno do ponto fixo:

$$g(x) = g(x^*) + (x - x^*)g'(x^*) + O\left((x - x^*)^2\right) \tag{3.56}$$

$$= x^* + (x - x^*)g'(x^*) + O\left((x - x^*)^2\right) \tag{3.57}$$

$$\approx x^* + (x - x^*)g'(x^*) \tag{3.58}$$

Substituindo na relação de recorrência, temos

$$x^{(n+1)} = g\left(x^{(n)}\right) \approx x^* + (x^{(n)} - x^*)g'(x^*) \tag{3.59}$$

Ou seja:

$$\left(x^{(n+1)} - x^*\right) \approx (x^{(n)} - x^*)g'(x^*) \tag{3.60}$$

Tomando módulos, temos:

$$\underbrace{\left|x^{(n+1)} - x^*\right|}_{\epsilon_{n+1}} \approx \underbrace{\left|x^{(n)} - x^*\right|}_{\epsilon_n} \left|g'(x^*)\right|, \tag{3.61}$$

onde $\epsilon_n = \left|x^{(n)} - x^*\right|$.

Observação 3.3.3. Da análise acima, concluímos:

- Se $|g'(x^*)| < 1$, então, a distância de $x^{(n)}$ até o ponto fixo x^* está diminuindo a cada passo.
- Se $|g'(x^*)| > 1$, então, a distância de $x^{(n)}$ até o ponto fixo x^* está aumentando a cada passo.
- Se $|g'(x^*)| = 1$, então, nossa aproximação de primeira ordem não é suficiente para compreender o comportamento da sequência.

3.3.4 Erro absoluto e tolerância

Na prática, quando se aplica uma iteração como esta, não se conhece de antemão o valor do ponto fixo x^*. Assim, o erro $\epsilon_n = \left|x^{(n)} - x^*\right|$ precisa ser estimado com base nos valores calculados $x^{(n)}$. Uma abordagem frequente é analisar a evolução da diferença entre dois elementos da sequência:

$$\Delta_n = \left|x^{(n+1)} - x^{(n)}\right| \tag{3.62}$$

A pergunta natural é: Será que o erro $\epsilon_n = \left|x^{(n)} - x^*\right|$ será pequeno quando $\Delta_n = \left|x^{(n+1)} - x^{(n)}\right|$ for pequeno?

Para responder a esta pergunta, observamos que

$$x^* = \lim_{n\to\infty} x^{(n)} \tag{3.63}$$

portanto:

$$\begin{aligned} x^* - x^{(N)} &= \left(x^{(N+1)} - x^{(N)}\right) + \left(x^{(N+2)} - x^{(N+1)}\right) + \left(x^{(N+3)} - x^{(N+2)}\right) + \quad (3.64)\\ &= \sum_{k=0}^{\infty} \left(x^{(N+k+1)} - x^{(N+k)}\right) \quad (3.65) \end{aligned}$$

Usamos também as expressões:

$$\begin{aligned} x^{(n+1)} &\approx x^* + (x^{(n)} - x^*)g'(x^*) \quad (3.66)\\ x^{(n)} &\approx x^* + (x^{(n-1)} - x^*)g'(x^*) \quad (3.67) \end{aligned}$$

Subtraindo uma da outra, temos:

$$x^{(n+1)} - x^{(n)} \approx (x^{(n)} - x^{(n-1)})g'(x^*) \tag{3.68}$$

Portanto:

$$x^{(N+k+1)} - x^{(N+k)} \approx (x^{(N+1)} - x^{(N)})\left(g'(x^*)\right)^k \tag{3.69}$$

E temos:

$$x^* - x^{(N)} = \sum_{k=0}^{\infty}\left(x^{(N+k+1)} - x^{(N+k)}\right) \tag{3.70}$$

$$\approx \sum_{k=0}^{\infty}(x^{(N+1)} - x^{(N)})\left(g'(x^*)\right)^k \tag{3.71}$$

$$= (x^{(N+1)} - x^{(N)})\frac{1}{1-g'(x^*)}, \quad |g'(x^*)| < 1 \tag{3.72}$$

Tomando módulo, temos:

$$\left|x^* - x^{(N)}\right| \approx \left|x^{(N+1)} - x^{(N)}\right|\frac{1}{1-g'(x^*)} \tag{3.73}$$

$$\epsilon_N \approx \frac{\Delta_N}{1-g'(x^*)} \tag{3.74}$$

Observação 3.3.4. Tendo em mente a relação $x^{(n+1)} - x^{(n)} \approx (x^{(n)} - x^{(n-1)})g'(x^*)$, concluímos:

- Quando $g'(x^*) < 0$, o esquema é alternante, isto é, o sinal do erro se altera a cada passo. O erro ϵ_N pode ser estimado diretamente da diferença Δ_N, pois o denominador $1 - g'(x^*) > 1$.
- Quando $0 < g'(x^*) < 1$, o esquema é monótono e $\frac{1}{1-g'(x^*)} > 1$, pelo que o erro ϵ_N é maior que a diferença Δ_N. A relação será tão mais importante quando mais próximo da unidade for $g'(x^*)$, ou seja, quando mais lenta for a convergência. Para estimar o erro em função da diferença Δ_N, observamos que $g'(x^*) \approx \frac{x^{(n+1)} - x^{(n)}}{x^{(n)} - x^{(n-1)}}$ e

  $$|g'(x^*)| \approx \frac{\Delta_n}{\Delta_{n-1}} \tag{3.75}$$

  e portanto

  $$\epsilon_N \approx \frac{\Delta_N}{1 - \frac{\Delta_n}{\Delta_{n-1}}}. \tag{3.76}$$

Exercícios

E 3.3.1. Resolver a equação $e^x = x + 2$ é equivalente a calcular os pontos fixos da função $g(x) = e^x - 2$ (veja o Exemplo 3.3.1). Use a iteração do ponto fixo $x^{(n+1)} = g(x^n)$ com $x^{(1)} = -1{,}8$ para obter uma aproximação de uma das soluções da equação dada com 8 dígitos significativos.

E 3.3.2. Mostre que a equação:

$$\cos(x) = x \tag{3.77}$$

possui uma única solução no intervalo $[0, 1]$. Use a iteração do ponto fixo e encontre uma aproximação para esta solução com 4 dígitos significativos.

E 3.3.3. Mostre que a equação $xe^x = 10$ é equivalente às seguintes equações:

$$x = \ln\left(\frac{10}{x}\right) \quad \text{e} \quad x = 10e^{-x}. \tag{3.78}$$

Destas, considere as seguintes iterações de ponto fixo:

a) $x^{(n+1)} = \ln\left(\dfrac{10}{x^{(n)}}\right)$

b) $x^{(n+1)} = 10e^{-x^{(n)}}$

Tomando $x^{(1)} = 1$, verifique se estas sequências são convergentes.

E 3.3.4. Verifique (analiticamente) que a única solução real da equação:

$$xe^x = 10 \tag{3.92}$$

é ponto fixo das seguintes funções:

a) $g(x) = \ln\left(\frac{10}{x}\right)$

b) $g(x) = x - \frac{xe^x - 10}{15}$

c) $g(x) = x - \frac{xe^x - 10}{10 + e^x}$

Implemente o processo iterativo $x^{(n+1)} = g(x^{(n)})$ para $n \geq 0$ e compare o comportamento. Discuta os resultados com base na teoria estudada.

E 3.3.5. Verifique (analiticamente) que a única solução real da equação:

$$\cos(x) = x \tag{3.93}$$

é ponto fixo das seguintes funções:

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

b) $g(x) = 0{,}4x + 0{,}6\cos(x)$

c) $g(x) = x + \frac{\cos(x)-x}{1+\text{sen}\,(x)}$

Implemente o processo iterativo $x^{(n+1)} = g(x^{(n)})$ para $n \geq 0$ e compare o comportamento. Discuta os resultados com base na teoria estudada.

E 3.3.6. Encontre a solução de cada equação com erro absoluto inferior a 10^{-6}.

a) $e^x = x + 2$ no intervalo $(-2{,}0)$.

b) $x^3 + 5x^2 - 12 = 0$ no intervalo $(1{,}2)$.

c) $\sqrt{x} = \cos(x)$ no intervalo $(0{,}1)$.

E 3.3.7. Encontre numericamente as três primeiras raízes positivas da equação dada por:

$$\cos(x) = \frac{x}{10 + x^2} \tag{3.94}$$

com erro absoluto inferior a 10^{-6}.

E 3.3.8. Considere os seguintes processos iterativos:

$$a \begin{cases} x^{(n+1)} &= \cos(x^{(n)}) \\ x^{(1)} &= .5 \end{cases}$$

$$\text{e} \tag{3.95}$$

$$b \begin{cases} x^{(n+1)} &= .4x^{(n)} + .6\cos(x^{(n)}) \\ x^{(1)} &= .5 \end{cases}$$

Use o teorema do ponto fixo para verificar que cada um desses processos converge para a solução da equação x^* de $\cos(x) = x$. Observe o comportamento numérico dessas sequências. Qual estabiliza mais rápido com cinco casas decimais? Discuta.

Dica: Verifique que $\cos([0.5{,}1]) \subseteq [0.5{,}1]$ e depois a mesma identidade para a função $f(x) = 0{,}4x + 0{,}6\cos(x)$.

E 3.3.9. Use o teorema do ponto fixo aplicado a um intervalo adequado para mostrar que a função $g(x) = \ln(100 - x)$ possui um ponto fixo estável.

E 3.3.10. (Fluidos) Na hidráulica, o fator de atrito de Darcy é dado pela implicitamente pela equação de Colebrook-White:

$$\frac{1}{\sqrt{f}} = -2\log_{10}\left(\frac{\varepsilon}{14.8R_h} + \frac{2.51}{\mathrm{Re}\sqrt{f}}\right) \tag{3.96}$$

onde f é o fator de atrito, ε é a rugosidade do tubo em metros, R_h é o raio hidráulico em metros e Re é o número de Reynolds. Considere $\varepsilon = 2mm$, $R_h = 5cm$ e $Re = 10000$ e obtenha o valor de f pela iteração:

$$x^{(n+1)} = -2\log_{10}\left(\frac{\varepsilon}{14.8R_h} + \frac{2.51x^{(n)}}{\mathrm{Re}}\right) \tag{3.97}$$

E 3.3.11. Encontre uma solução aproximada para a equação algébrica

$$180 - 100x = 0.052\,\mathrm{senh}^{\,-1}(10^{13}x) \tag{3.98}$$

com erro absoluto inferior a 10^{-3} usando um método iterativo. Estime o erro associado ao valor de $v = 180 - 100x = 0.052\,\mathrm{senh}^{\,-1}(10^{13}x)$, usando cada uma dessas expressões. Discuta sucintamente o resultado obtido. Dica: Este caso é semelhante ao Problema 3.2.8.

E 3.3.12. Considere que x_n satisfaz a seguinte relação de recorrência:

$$x_{n+1} = x_n - \beta\,(x_n - x^*) \tag{3.99}$$

onde β e x^* são constantes. Prove que

$$x_n - x^* = (1-\beta)^{n-1}(x_1 - x^*). \tag{3.100}$$

Conclua que $x_n \to x^*$ quando $|1-\beta| < 1$.

E 3.3.13. (Convergência lenta) Considere o seguinte esquema iterativo:

$$x^{(n+1)} = x_n + q^n, \tag{3.101}$$
$$x^{(0)} = 0, \tag{3.102}$$

onde $q = 1 - 10^{-6}$.

a) Calcule o limite

$$x_\infty = \lim_{n\to\infty} x^{(n)} \tag{3.103}$$

analiticamente.

b) Considere que o problema de obter o limite da sequência numericamente usando como critério de parada que $|x^{(n+1)} - x^{(n)}| < 10^{-5}$. Qual o valor é produzido pelo esquema numérico? Qual o desvio entre o valor obtido pelo esquema numérico e o valor do limite obtido no item a? Discuta. (Dica: Você não deve implementar o esquema iterativo, obtendo o valor de $x^{(n)}$ analiticamente)

c) Qual deve ser a tolerância especificada para obter o resultado com erro relativo inferior a 10^{-2}?

E 3.3.14. (Convergência sublinear) Considere o seguinte esquema iterativo:

$$x^{(n+1)} = x^{(n)} - [x^{(n)}]^3,\ x^{(n)} \geq 0 \tag{3.104}$$

com $x^{(0)} = 10^{-2}$. Prove que $\{x^{(n)}\}$ é sequência de número reais positivos convergindo para zero. Verifique que são necessários mais de mil passos para que $x^{(n)}$ se torne menor que $0.9x^{(0)}$.

E 3.3.15. (Taxa de convergência)

a) Use o teorema do ponto fixo para mostrar que a função $g(x) = 1 - \operatorname{sen}(x)$ possui um único ponto fixo estável o intervalo $[\frac{1}{10},1]$. Construa um método iterativo $x^{(n+1)} = g(x^{(n)})$ para encontrar esse ponto fixo. Use o computador para encontrar o valor numérico do ponto fixo.

b) Verifique que função $\psi(x) = \frac{1}{2}[x + 1 - \operatorname{sen}(x)]$ possui um ponto fixo x^* que também é o ponto fixo da função g do item a. Use o computador para encontrar o valor numérico do ponto fixo através da iteração $x^{(n+1)} = \psi(x^{(n)})$. Qual método é mais rápido?

E 3.3.16. (Esquemas oscilantes)(*Esquemas oscilantes*)

a) Considere a função $g(x)$ e a função composta $\psi(x) = g \circ g = g(g(x))$. Verifique todo ponto fixo de g também é ponto fixo de ψ.

b) Considere a função

$$g(x) = 10\exp(-x) \tag{3.105}$$

e a função composta $\psi(x) = g \circ g = g(g(x))$. Mostre que ψ possui dois pontos fixos que não são pontos fixos de g.

c) No problema anterior, o que acontece quando o processo iterativo $x^{(n+1)} = g(x^{(n)})$ é inicializado com um ponto fixo de ψ que não é ponto fixo de g?

E 3.3.17. (Aceleração de convergência - introdução ao método de Newton) Mostre que se $f(x)$ possui uma raiz x^* então a x^* é um ponto fixo de $\phi(x) = x + \gamma(x)f(x)$. Encontre uma condição em $\gamma(x)$ para que o ponto fixo x^* de ϕ seja estável. Encontre uma condição em $\gamma(x)$ para que $\phi'(x^*) = 0$.

E 3.3.18. (Aceleração de convergência - introdução ao método de Newton) Considere que $x^{(n)}$ satisfaz a seguinte relação de recorrência:

$$x^{(n+1)} = x^{(n)} - \gamma f(x^{(n)}) \tag{3.106}$$

onde γ é uma constante. Suponha que $f(x)$ possui um zero em x^*. Aproxime a função $f(x)$ em torno de x^* por

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + O\left((x - x^*)^2\right). \tag{3.107}$$

Em vista do problema anterior, qual valor de γ você escolheria para que a sequência $x^{(n)}$ convirja rapidamente para x^*.

E 3.3.19. Considere o problema da Questão 3.2.8 e dois seguintes esquemas iterativos.

$$A\begin{cases} I^{(n+1)} = \frac{1}{R}\left[V - v_t \ln\left(1 + \frac{I^{(n)}}{I_R}\right)\right], n > 0 \\ I^{(0)} = 0 \end{cases}$$

$$\text{e} \tag{3.108}$$

$$B\begin{cases} I^{(n+1)} = I_R\left[\exp\left(\frac{V - RI^{(n)}}{v_t}\right) - 1\right], n > 0 \\ I^{(0)} = 0 \end{cases}$$

Verifique numericamente que apenas o processo A é convergente para a, b e c; enquanto apenas o processo B é convergente para os outros itens.

3.4 Método de Newton-Raphson

Nesta seção, apresentamos o **método de Newton-Raphson**[5][6] para calcular o zero de funções reais de uma variável real.

Consideramos que x^* seja um zero de uma dada função $y = f(x)$ continuamente diferenciável, isto é, $f(x^*) = 0$. A fim de usar a iteração do ponto fixo, observamos que, equivalentemente, x^* é um ponto fixo da função:

$$g(x) = x + \alpha(x)f(x), \quad \alpha(x) \neq 0, \tag{3.109}$$

[5]Joseph Raphson, 1648 - 1715, matemático inglês.
[6]Também chamado apenas de método de Newton.

onde $\alpha(x)$ é uma função arbitrária, a qual escolheremos de forma que a iteração do ponto fixo tenha ótima taxa de convergência.

Do **teorema do ponto fixo**, a taxa de convergência é dada em função do valor absoluto da derivada de $g(x)$. Calculando a derivada temos:

$$g'(x) = 1 + \alpha(x)f'(x) + \alpha'(x)f(x). \tag{3.110}$$

No ponto $x = x^*$, temos:

$$g'(x^*) = 1 + \alpha(x^*)f'(x^*) + \alpha'(x^*)f(x^*). \tag{3.111}$$

Como $f(x^*) = 0$, temos:

$$g'(x^*) = 1 + \alpha(x^*)f'(x^*). \tag{3.112}$$

Sabemos que o processo iterativo converge tão mais rápido quanto menor for $|g'(x)|$ nas vizinhanças de x^*. Isto nos leva a escolher:

$$g'(x^*) = 0, \tag{3.113}$$

e, então, temos:

$$\alpha(x^*) = -\frac{1}{f'(x^*)}, \tag{3.114}$$

se $f'(x^*) \neq 0$.

A discussão acima nos motiva a introduzir o método de Newton, cujas iterações são dada por:

$$x^{(n+1)} = x^{(n)} - \frac{f\left(x^{(n)}\right)}{f'\left(x^{(n)}\right)}, \quad n \geq 1, \tag{3.115}$$

sendo $x^{(1)}$ uma aproximação inicial dada.

3.4.1 Interpretação geométrica

Seja uma dada função $f(x)$ conforme na Figura 3.6. Para tanto, escolhemos uma aproximação inicial $x^{(1)}$ e computamos:

$$x^{(2)} = x^{(1)} - \frac{f(x^{(1)})}{f'(x^{(1)})}. \tag{3.116}$$

Geometricamente, o ponto $x^{(2)}$ é a interseção da reta tangente ao gráfico da função $f(x)$ no ponto $x = x^{(1)}$ com o eixo das abscissas. Com efeito, a equação desta reta é:

$$y = f'(x^{(1)})(x - x^{(1)}) + f(x^{(1)}). \tag{3.117}$$

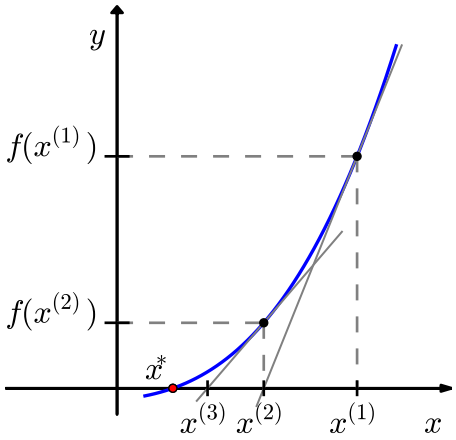


Figure 3.6: Interpretação do método de Newton.

Assim, a interseção desta reta com o eixo das abscissas ($y = 0$) ocorre quando:

$$f'(x^{(1)})(x - x^{(1)}) + f(x^{(1)}) = 0 \Rightarrow x = x^{(1)} - \frac{f(x^{(1)})}{f'(x^{(1)})}. \tag{3.118}$$

Ou seja, dada aproximação $x^{(n)}$, a próxima aproximação $x^{(n+1)}$ é o ponto de interseção entre o eixo das abscissas e a reta tangente ao gráfico da função no ponto $x = x^{(n)}$. Observe a Figura 3.6.

3.4.2 Análise de convergência

Seja $y = f(x)$ uma função com derivadas primeira e segunda contínuas tal que $f(x^*) = 0$ e $f'(x^*) \neq 0$. Seja também a função $g(x)$ definida como:

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{3.119}$$

Expandindo em série de Taylor em torno de $x = x^*$, obtemos:

$$g(x) = g(x^*) + g'(x^*)(x - x^*) + \frac{g''(x^*)}{2}(x - x^*)^2 + O\left((x - x^*)^3\right). \tag{3.120}$$

Observamos que:

$$g(x^*) = x^* \tag{3.121}$$

$$g'(x^*) = 1 - \frac{f'(x^*)f'(x^*) - f(x^*)f''(x^*)}{(f'(x^*))^2} = 0 \tag{3.122}$$

Portanto:

$$g(x) = x^* + \frac{g''(x^*)}{2}(x - x^*)^2 + O\left((x - x^*)^3\right) \tag{3.123}$$

Com isso, temos:

$$x^{(n+1)} = g(x^{(n)}) = x^* + \frac{g''(x^*)}{2}(x^{(n)} - x^*)^2 + O\left((x^{(n)} - x^*)^3\right), \tag{3.124}$$

ou seja, para n suficientemente grande,

$$\left|x^{(n+1)} - x^*\right| \leq C\left|x^{(n)} - x^*\right|^2, \tag{3.125}$$

com constante $C = |g''(x^*)/2|$. Isto mostra que o método de Newton tem **taxa de convergência quadrática**. Mais precisamente, temos o seguinte teorema.

Teorema 3.4.1 (Método de Newton)**.** *Sejam $f \in C^2([a,b])$ com $x^* \in (a,b)$ tal que $f(x^*) = 0$ e:*

$$m := \min_{x\in[a,b]} |f'(x)| > 0 \quad e \quad M := \max_{x\in[a,b]} |f''(x)|. \tag{3.126}$$

Escolhendo $\rho > 0$ tal que:

$$q := \frac{M}{2m}\rho < 1, \tag{3.127}$$

*definimos a **bacia de atração** do método de Newton pelo conjunto:*

$$K_\rho(x^*) := \{x \in \mathbb{R};\ |x - x^*| \leq \rho\} \subset [a,b]. \tag{3.128}$$

Então, para qualquer $x^{(1)} \in K_\rho(x^*)$ a iteração do método de Newton:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}, \tag{3.129}$$

*fornece uma sequência $x^{(n)}$ que converge para x^*, isto é, $x^{(n)} \to x^*$ quando $n \to \infty$. Além disso, temos a seguinte estimativa de erro **a priori**:*

$$|x^{(n)} - x^*| \leq \frac{2m}{M} q^{(2^{n-1})}, \quad n \geq 2, \tag{3.130}$$

*e a seguinte estimativa de erro **a posteriori**:*

$$|x^{(n)} - x^*| \leq \frac{M}{2m}|x^{(n)} - x^{(n-1)}|^2, \quad n \geq 2. \tag{3.131}$$

Proof. Para $n \in \mathbb{N}$, $n \geq 2$, temos:

$$x^{n+1} - x^* = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})} - x^* = -\frac{1}{f'(x^{(n)})}\left[f(x^{(n)}) + (x^* - x^{(n)})f'(x^{(n)})\right]. \tag{3.132}$$

Agora, para estimar o lado direito desta equação, usamos o polinômio de Taylor de grau 1 da função $f(x)$ em torno de $x = x^{(n)}$, isto é:

$$f(x^*) = f(x^{(n)}) + (x^* - x^{(n)})f'(x^{(n)}) + \int_{x^{(n)}}^{x^*} f''(t)(x^* - t)\, dt. \tag{3.133}$$

Pela mudança de variável $t = x^{(n)} + s(x^* - x^{(n)})$, observamos que o resto deste polinômio de Taylor na forma integral é igual a:

$$R(x^*,x^{(n)}) := (x^* - x^{(n)})^2 \int_0^1 f''\left(x^{(n)} + s(x^* - x^{(n)})\right)(1 - s)\, ds. \tag{3.134}$$

Assim, da cota da segunda derivada de $f(x)$, temos:

$$|R(x^*,x^{(n)})| \leq M|x^* - x^{(n)}|^2 \int_0^1 (1 - s)\, ds = \frac{M}{2}|x^* - x^{(n)}|^2. \tag{3.135}$$

Se $x^{(n)} \in K_\rho(x^*)$, então de (3.132) e (3.135) temos:

$$|x^{(n+1)} - x^*| \leq \frac{M}{2m}|x^{(n)} - x^*|^2 \leq \frac{M}{2m}\rho^2 < \rho. \tag{3.136}$$

Isto mostra que se $x^{(n)} \in K_\rho(x^*)$, então $x^{(n+1)} \in K_\rho(x^*)$, isto é, $x^{(n)} \in K_\rho(x^*)$ para todo $n \in \mathbb{R}$.

Agora, obtemos a estimativa **a priori** de (3.4.2), pois:

$$|x^{(n)} - x^*| \leq \frac{2m}{M}\left(\frac{M}{2m}|x^{(n-1)} - x^*|\right)^2 \leq \cdots \leq \frac{2m}{M}\left(\frac{M}{2m}|x^{(1)} - x^*|\right)^{2^{n-1}}. \tag{3.137}$$

Logo:

$$|x^{(n)} - x^*| \leq \frac{2m}{M}q^{2^{n-1}}, \tag{3.138}$$

donde também vemos que $x^{(n)} \to x^*$ quando $n \to \infty$, pois $q < 1$.

Por fim, para provarmos a estimativa **a posteriori** tomamos a seguinte expansão em polinômio de Taylor:

$$f(x^{(n)}) = f(x^{(n-1)}) + (x^{(n)} - x^{(n-1)})f'(x^{(n-1)}) + R(x^{(n)},x^{(n-1)}). \tag{3.139}$$

Aqui, temos:

$$f(x^{(n-1)}) + (x^{(n)} - x^{(n-1)})f'(x^{(n-1)}) = 0 \tag{3.140}$$

e, então, conforme acima:

$$|f(x^{(n)})| = |R(x^{(n)},x^{(n-1)})| \leq \frac{M}{2}|x^{(n)} - x^{(n-1)}|^2. \tag{3.141}$$

Com isso e do teorema do valor médio, concluímos:

$$|x^{(n)} - x^*| \leq \frac{1}{m}|f(x^{(n)}) - f(x^*)| \leq \frac{M}{2m}|x^{(n)} - x^{(n-1)}|^2. \tag{3.142}$$

□

Exemplo 3.4.1. Estime o raio ρ da bacia de atração $K_\rho(x^*)$ para a função $f(x) = \cos(x) - x$ restrita ao intervalo $[0, \pi/2]$.

Solução. O raio da bacia de atração é tal que:

$$\rho < \frac{2m}{M} \tag{3.143}$$

onde $m := \min|f'(x)|$ e $M := \max|f''(x)|$ com o mínimo e o máximo tomados em um intervalo $[a, b]$ que contenha o zero da função $f(x)$. Aqui, por exemplo, podemos tomar $[a, b] = [0, \pi/2]$. Como, neste caso, $f'(x) = -\operatorname{sen}(x) - 1$, temos que $m = 1$. Também, como $f''(x) = -\cos x$, temos $M = 1$. Assim, concluímos que $\rho < 2$ (lembrando que $K_\rho(x^*) \subset [0, \pi/2]$). Ou seja, neste caso as iterações de Newton convergem para o zero de $f(x)$ para qualquer escolha da aproximação inicial $x^{(1)} \in [0, \pi/2]$. ◊

Exercícios

E 3.4.1. Encontre a raiz positiva da função $f(x) = \cos(x) - x^2$ pelo método de Newton inicializando-o com $x^{(0)} = 1$. Realize a iteração até obter estabilidade no *quinto* dígito significativo.

E 3.4.2. Considere o problema de calcular as soluções positivas da equação:

$$\operatorname{tg}(x) = 2x^2. \tag{3.144}$$

a) Use o método gráfico para isolar as duas primeiras raízes positivas em pequenos intervalos. Use a teoria para argumentar quanto à existência e unicidade das raízes dentro intervalos escolhidos.

b) Calcule cada uma das raízes pelo método de Newton com oito dígitos significativos e discuta a convergência.

E 3.4.3. Considere a equação

$$e^{-x^2} = x \tag{3.145}$$

trace o gráfico com auxílio do computador e verifique que ela possui uma raiz positiva. Encontre uma aproximação para esta raiz pelo gráfico e use este valor para inicializar o método de Newton e obtenha uma aproximação para a raiz com 8 dígitos significativos. (Use o comando `format('v',16)` para alterar a visualização no `Scilab`.)

E 3.4.4. Isole e encontre as cinco primeiras raízes positivas da equação com 6 dígitos corretos através de traçado de gráfico e do método de Newton.

$$\cos(10x) = e^{-x}. \tag{3.146}$$

Dica: a primeira raiz positiva está no intervalo $(0, 0{,}02)$. Fique atento.

E 3.4.5. Encontre as raízes do polinômio $f(x) = x^4 - 4x^2 + 4$ através do método de Newton. O que você observa em relação ao erro obtido? Compare com a situação do Problema 3.2.4.

E 3.4.6. Encontre as raízes reais do polinômio $f(x) = \frac{x^5}{100} + x^4 + 3x + 1$ isolando-as pelo método do gráfico e depois usando o método de Newton. Expresse a solução com 7 dígitos significativos.

E 3.4.7. Considere o método de Newton aplicado para encontrar a raiz de $f(x) = x^3 - 2x + 2$. O que acontece quando $x^{(0)} = 0$? Escolha um valor adequado para inicializar o método e obter a única raiz real desta equação.

E 3.4.8. Justifique a construção do processo iterativo do método de Newton através do conceito de estabilidade de ponto fixo e convergência do método da iteração. Dica: Considere os problemas 3.3.17 e 3.3.18.

E 3.4.9. Entenda a interpretação geométrica ao método de Newton. Encontre uma valor para iniciar o método de Newton aplicado ao problema $f(x) = xe^{-x} = 0$ tal que o esquema iterativo divirja.

E 3.4.10. (Computação) Aplique o método de Newton à função $f(x) = \frac{1}{x} - A$ e construa um esquema computacional para calcular a inversa de A com base em operações de multiplicação e soma/subtração.

E 3.4.11. (Computação) Aplique o método de Newton à função $f(x) = x^n - A$ e construa um esquema computacional para calcular $\sqrt[n]{A}$ para $A > 0$ com base em operações de multiplicação e soma/subtração.

E 3.4.12. (Computação) Aplique o método de Newton à função $f(x) = \frac{1}{x^2} - A$ e construa um esquema computacional para calcular $\frac{1}{\sqrt{A}}$ para $A > 0$ com base em operações de multiplicação e soma/subtração.

E 3.4.13. Considere a função dada por

$$\psi(x) \quad = \quad \ln\left(15 - \ln(x)\right) \tag{3.155}$$

definida para $x \in (0,e^{15})$

a) Use o teorema do ponto fixo para provar que se $x^{(0)}$ pertence ao intervalo [1,3], então a sequência dada iterativamente por

$$x^{(n+1)} = \psi(x^{(n)}), n \geq 0 \tag{3.156}$$

converge para o único ponto fixo, x^*, de ψ. Construa a iteração $x^{(n+1)} = \psi(x^{(n)})$ e obtenha numericamente o valor do ponto fixo x^*. Expresse a resposta com 5 algarismos significativos corretos.

b) Construa a iteração do método de Newton para encontrar x^*, explicitando a relação de recorrência e iniciando com $x_0 = 2$. Use o computador para obter a raiz e expresse a resposta com oito dígitos significativos corretos.

3.5 Método das secantes

O **método das secantes** é uma variação do método de Newton, evitando a necessidade de conhecer-se a derivada analítica de $f(x)$. Dada uma função $f(x)$, a ideia é aproximar sua derivada pela razão fundamental:

$$f'(x) \approx \frac{f(x) - f(x_0)}{x - x_0}, \quad x \approx x_0. \tag{3.157}$$

Mais precisamente, o método de Newton é uma iteração de ponto fixo da forma:

$$x^{(n+1)} = x^{(n)} - \alpha(x^{(n)}) f(x^{(n)}), \quad n \geq 1, \tag{3.158}$$

onde $x^{(1)}$ é uma aproximação inicial dada e $\alpha(x^{(n)}) = 1/f'(x^{(n)})$. Usando a aproximação da derivada acima, com $x = x^{(n)}$ e $x_0 = x^{(n-1)}$, temos:

$$\alpha(x^{(n)}) = \frac{1}{f'(x^{(n)})} \approx \frac{x^{(n)} - x^{(n-1)}}{f(x^{(n)}) - f(x^{(n-1)})}. \tag{3.159}$$

Isto nos motiva a introduzir a **iteração do método das secantes** dada por:

$$x^{(n+1)} = x^{(n)} - f(x^{(n)})\frac{x^{(n)} - x^{(n-1)}}{f(x^{(n)}) - f(x^{(n-1)})}, \quad n \geq 2. \tag{3.160}$$

Observe que para inicializarmos a iteração acima precisamos de duas aproximações iniciais, a saber, $x^{(1)}$ e $x^{(2)}$. Maneiras apropriadas de escolher estas aproximações podem ser inferidas da interpretação geométrica do método.

Exemplo 3.5.1. Encontre as raízes de $f(x) = \cos(x) - x$.

Solução. Da inspeção do gráfico das funções $y = \cos(x)$ e $y = x$, sabemos que esta equação possui uma raiz em torno de $x = 0{,}8$. Iniciamos o método com $x_0 = 0{,}7$ e $x_1 = 0{,}8$.

$x^{(n-1)}$	$x^{(n)}$	m	$x^{(n+1)}$
0,7	0,8	$\frac{f(0,8)-f(0,7)}{0,8-0,7} = -1{,}6813548$	$0{,}8 - \frac{f(0,8)}{-1,6813548} = 0{,}7385654$
0,8	0,7385654	−1,6955107	0,7390784
0,7385654	0,7390784	−1,6734174	0,7390851
0,7390784	0,7390851	−1,6736095	0,7390851

◊

3.5.1 Interpretação geométrica

Enquanto, o método de Newton está relacionado às retas tangentes ao gráfico da função objetivo $f(x)$, o método das secantes, como o próprio nome indica, está relacionado às retas secantes.

Sejam $f(x)$ e as aproximações $x^{(1)}$ e $x^{(2)}$ do zero x^* desta função (veja Figura 3.7). A iteração do método das secantes fornece:

$$x^{(3)} = x^{(2)} - f(x^{(2)})\frac{x^{(2)} - x^{(1)}}{f(x^{(2)}) - f(x^{(1)})}. \tag{3.161}$$

De fato, $x^{(3)}$ é o ponto de interseção da reta secante ao gráfico de $f(x)$ pelos pontos $x^{(1)}$ e $x^{(2)}$ com o eixo das abscissas. Com efeito, a equação desta reta secante é:

$$y = \frac{f(x^{(2)}) - f(x^{(1)})}{x^{(2)} - x^{(1)}}(x - x^{(2)}) + f(x^{(2)}). \tag{3.162}$$

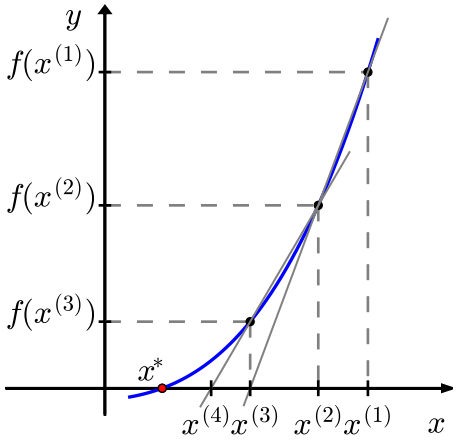


Figure 3.7: Método das secantes.

Esta reta intercepta o eixo das abscissas no ponto x tal que $y = 0$, isto é:

$$\frac{f(x^{(2)}) - f(x^{(1)})}{x^{(2)} - x^{(1)}}(x - x^{(2)}) + f(x^{(2)}) = 0 \Rightarrow x = x^{(2)} - f(x^{(2)})\frac{x^{(2)} - x^{(1)}}{f(x^{(2)}) - f(x^{(1)})}. \tag{3.163}$$

3.5.2 Análise de convergência

Uma análise assintótica semelhante àquela feita para o método de Newton na subseção 3.4.2 nos indica que, para uma função $f(x)$ duas vezes diferenciável, as iterações do método da secante satisfazem:

$$|x^{(n+1)} - x^*| \approx C|x^{(n)} - x^*||x^{(n-1)} - x^*|, \tag{3.164}$$

para aproximações iniciais suficientemente próximas de x^*, onde $f(x^*) = 0$. Além disso, veremos que:

$$|x^{(n+1)} - x^*| \leq C|x^{(n)} - x^*|^p, \;\; p = \frac{\sqrt{5}+1}{2} \approx 1{,}618 \tag{3.165}$$

sob certas condições. Ou seja, o método das secantes tem **taxa de convergência superlinear**.

Teorema 3.5.1 (Método das secantes)**.** *Seja $f \in C^2([a,b])$ uma função com $x^* \in (a,b)$ tal que $f(x^*) = 0$. Sejam, também:*

$$m := \min_{x\in[a,b]} |f'(x)| > 0 \quad e \quad M := \max_{x\in[a,b]} |f''(x)| < \infty. \tag{3.166}$$

Além disso, seja $\rho > 0$ tal que:

$$q := \frac{M}{2m}\rho < 1, \quad K_\rho(x^*) := \{x \in \mathbb{R};\ |x - x^*| \leq \rho\} \subset [a, b]. \tag{3.167}$$

*Então, para aproximações iniciais $x^{(1)}, x^{(2)} \in K_\rho(x^*)$, com $x^{(1)} \neq x^{(2)}$, temos que as iterações do método das secantes $x^{(n)} \in K_\rho(x^*)$, $n \geq 1$, e $x^{(n)} \to x^*$, quando $n \to \infty$. Além disso, vale a seguinte estimativa de convergência **a priori**:*

$$|x^{(n)} - x^*| \leq \frac{2m}{M} q^{\gamma_{n-1}}, \quad n \geq 1, \tag{3.168}$$

*onde $\{\gamma_n\}_{n\in\mathbb{N}}$ é a sequência de Fibonacci[7][8], bem como vale a estimativa **a posteriori**:*

$$|x^{(n)} - x^*| \leq \frac{M}{2m}|x^{(n)} - x^{(n-1)}||x^{(n-1)} - x^{(n-2)}|, \quad n \geq 3. \tag{3.169}$$

Proof. Sejam $n \in \mathbb{N}$ com $n \geq 2$ e $x^{(n)}, x^{(n-1)} \in K_\rho(x^*)$, tal que $x^{(n)} \neq x^{(n-1)}$, $x^{(n)} \neq x^*$ e $x^{(n-1)} \neq x^*$. Seja, também:

$$g(x^{(n)},x^{(n-1)}) := x^{(n)} - f(x^{(n)})\frac{x^{(n)} - x^{(n-1)}}{f(x^{(n)}) - f(x^{(n-1)})}. \tag{3.170}$$

Com isso, temos:

$$g(x^{(n)},x^{(n-1)}) - x^* = x^{(n)} - f(x^{(n)})\frac{x^{(n)} - x^{(n-1)}}{f(x^{(n)}) - f(x^{(n-1)})} - x^* \tag{3.171}$$

$$= \frac{x^{(n)} - x^{(n-1)}}{f(x^{(n)}) - f(x^{(n-1)})}\left\{(x^{(n)} - x^*)\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}\right. \tag{3.172}$$

$$\left. - f(x^{(n)}) + f(x^*)\right\}. \tag{3.173}$$

$$\tag{3.174}$$

Então, da cota assumida para primeira derivada de $f(x)$ e do teorema do valor médio, temos:

$$|g(x^{(n)},x^{(n-1)}) - x^*| \leq \frac{|x^{(n)} - x^*|}{m}\left|\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} - \frac{f(x^{(n)}) - f(x^*)}{x^{(n)} - x^*}\right|. \tag{3.175}$$

Agora, iremos estimar este último termo a direita. Para tanto, começamos observando que da expansão em polinômio de Taylor de ordem 0 da função $f(x)$ com resto na forma integral, temos:

$$\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} = -\int_0^1 \frac{d}{dr} f(x^{(n)} + r(x^{(n-1)} - x^{(n)}))\frac{dr}{x^{(n)} - x^{(n-1)}} \tag{3.176}$$

$$= \int_0^1 f'(x^{(n)} + r(x^{(n-1)} - x^{(n)}))\, dr \tag{3.177}$$

[7] Leonardo Fibonacci, c. 1170 - c. 1250, matemático italiano.

[8] A sequência de Fibonacci $\{\gamma_n\}_{n\in\mathbb{N}}$ é definida por $\gamma_0 = \gamma_1 = 1$ e $\gamma_{n+1} = \gamma_n - \gamma_{n-1}$, $n \geq 1$.

De forma análoga, temos:

$$\frac{f(x^{(n)}) - f(x^*)}{x^{(n)} - x^*} = \int_0^1 f'(x^{(n)} + r(x^* - x^{(n)}))\, dr \tag{3.178}$$

Logo, temos:

$$\begin{aligned} &\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} - \frac{f(x^{(n)}) - f(x^*)}{x^{(n)} - x^*} = \\ &\int_0^1 \left[f'(x^{(n)} + r(x^{(n-1)} - x^{(n)})) - f'(x^{(n)} + r(x^* - x^{(n)})) \right] dr. \end{aligned} \tag{3.179}$$

Agora, novamente temos:

$$\begin{aligned} &f'(x^{(n)} + r(x^{(n-1)} - x^{(n)})) - f'(x^{(n)} + r(x^* - x^{(n)})) \\ &= \int_0^r \frac{d}{ds} f'(x^{(n)} + r(x^{(n-1)} - x^{(n)}) + s(x^* - x^{(n-1)}))\, ds \\ &= \int_0^r f''(x^{(n)} + r(x^{(n-1)} - x^{(n)}) + s(x^* - x^{(n-1)}))\, ds (x^* - x^{(n-1)}). \end{aligned} \tag{3.180}$$

Retornando à Equação (3.179) e usando a cota para a segunda derivada, obtemos:

$$\left| \frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} - \frac{f(x^{(n)}) - f(x^*)}{x^{(n)} - x^*} \right| \leq \frac{M}{2} |x^{(n-1)} - x^*|. \tag{3.181}$$

Utilizando a Equação (3.175), obtemos:

$$|g(x^{(n)}, x^{(n-1)}) - x^*| \leq \frac{M}{2m} |x^{(n)} - x^*||x^{(n-1)} - x^*| \leq \frac{M}{2m} \rho^2 < \rho. \tag{3.182}$$

Portanto, concluímos que as iterações do método da secantes $x^{(n)}$ permanecem no conjunto $K_\rho(x^*)$, se começarem nele. Além disso, temos demonstrado que:

$$|x^{(n+1)} - x^*| \leq \frac{M}{2m} |x^{(n)} - x^*||x^{(n-1)} - x^*|. \tag{3.183}$$

Com isso, temos:

$$\rho_n := \frac{M}{2m} |x^{(n)} - x^*| \Rightarrow \rho_{n+1} \leq \rho_n \rho_{n-1}, \quad n \geq 2. \tag{3.184}$$

Como $\rho_1 \leq q$ e $\rho_2 \leq q$, temos $\rho_n \leq q^{\gamma_{n-1}}$, $n \geq 1$. Isto mostra a estimativa de convergência **a priori**:

$$|x^n - x^*| \leq \frac{2m}{M} q^{\gamma_{n-1}}. \tag{3.185}$$

Além disso, como $\gamma_n \to \infty$ quando $n \to \infty$ e $q < 1$, temos que as iterações do método das secantes $x^{(n)} \to x^*$ quando $n \to \infty$.

Por fim, mostramos a estimativa de convergência **a posteriori**. Para tanto, da cota assumida para a primeira derivada e do teorema do valor médio, temos, para $n \geq 3$:

$$\begin{aligned} |x^{(n)} - x^*| &\leq \frac{1}{m}|f(x^{(n)} - f(x^*)| && (3.186)\\ &= \frac{1}{m}\left|f(x^{(n-1)}) + (x^{(n)} - x^{(n-1)})\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}\right| && (3.187)\\ &= \frac{1}{m}\left|x^{(n)} - x^{(n-1)}\right|\left|\frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} + \frac{f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}\right| && (3.188) \end{aligned}$$

Agora, a iteração do método das secantes fornece:

$$x^{(n)} = x^{(n-1)} - f(x^{(n-1)})\frac{x^{(n-1)} - x^{(n-2)}}{f(x^{(n-1)}) - f(x^{(n-2)})} \tag{3.189}$$

e temos:

$$\frac{f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}} = -\frac{f(x^{(n-1)}) - f(x^{(n-2)})}{x^{(n-1)} - x^{(n-2)}}. \tag{3.190}$$

Portanto:

$$|x^{(n)} - x^*| \leq \frac{1}{m}|x^{(n)} - x^{(n-1)}|\left|\frac{f(x^{(n-1)}) - f(x^{(n)})}{x^{(n-1)} - x^{(n)}} - \frac{f(x^{(n-1)}) - f(x^{(n-2)})}{x^{(n-1)} - x^{(n-2)}}\right|. \tag{3.191}$$

Observamos que o último termo pode ser estimado como feito acima para o termo análogo na Inequação (3.175). Com isso, obtemos a estimativa desejada:

$$|x^{(n)} - x^*| \leq \frac{M}{2m}|x^{(n)} - x^{(n-1)}||x^{(n)} - x^{(n-2)}|. \tag{3.192}$$

□

Proposição 3.5.1 (Sequência de Fibonacci)**.** *A sequência de Fibonacci* $\{\gamma_n\}_{n\in\mathbb{N}}$ *é assintótica a* $\gamma_n \sim \lambda_1^{n+1}/\sqrt{5}$ *e:*

$$\lim_{n\to\infty}\frac{\gamma_{n+1}}{\gamma_n} = \lambda_1, \tag{3.193}$$

onde $\lambda_1 = (1+\sqrt{5})/2 \approx 1{,}618$ *é a porção áurea.*

Proof. A sequência de Fibonacci $\{\gamma_n\}_{n\in\mathbb{N}}$ é definida por $\gamma_0 = \gamma_1 = 1$ e $\gamma_{n+1} = \gamma_n + \gamma_{n-1}$, $n \geq 1$. Logo, satisfaz a seguinte equação de diferenças:

$$\gamma_{n+2} - \gamma_{n+1} - \gamma_n = 0, \quad n \in \mathbb{N}. \tag{3.194}$$

Tomando $\gamma_n = \lambda^n$, $\lambda \neq 0$ temos:

$$\lambda^n \left(\lambda^2 - \lambda - 1\right) = 0 \Rightarrow \lambda^2 - \lambda - 1 = 0 \Rightarrow \lambda_{1,2} = \frac{1 \pm \sqrt{5}}{2}. \tag{3.195}$$

Portanto, $\gamma_n = c_1\lambda_1^n + c_2\lambda_2^n$. Como $\gamma_0 = \gamma_1 = 1$, as constantes satisfazem:

$$\begin{aligned} &c_1 + c_2 = 1 \\ &c_1\lambda_1 + c_2\lambda_2 = 1 \end{aligned} \Rightarrow c_1 = \frac{1+\sqrt{5}}{2\sqrt{5}}, \quad c_2 = -\frac{1-\sqrt{5}}{2\sqrt{5}}. \tag{3.196}$$

Ou seja, obtemos a seguinte forma explícita para os números de Fibonacci:

$$\gamma_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right]. \tag{3.197}$$

Daí, segue imediatamente o enunciado. □

Observação 3.5.1. Sob as hipóteses do Teorema 3.5.1 e da Proposição 3.5.1, temos:

$$\lim_{n\to\infty} \frac{|x^{(n+1)} - x^*|}{|x^{(n)} - x^*|^{\lambda_1}} \leq \lim_{n\to\infty} \frac{M}{2m} |x^{(n)} - x^*|^{1-\lambda_1} |x^{(n-1)} - x^*| \tag{3.198}$$

$$\leq \lim_{n\to\infty} \left(\frac{2m}{M}\right)^{1-\lambda_1} q^{(2-\lambda_1)\lambda_1^n/\sqrt{5}} = 0. \tag{3.199}$$

Isto mostra que o método das secantes (nestas hipóteses) tem taxa de convergência superlinear ($\lambda_1 \approx 1{,}6$).

3.6 Critérios de parada

Quando usamos métodos iterativos precisamos determinar um critério de parada. A Tabela 3.4 indica critérios de parada usuais para os métodos que estudamos neste capítulo.

Observação 3.6.1. O erro na tabela sempre se refere ao erro absoluto esperado. Nos três últimos métodos, é comum que se exija como critério de parada que a condição seja satisfeita por alguns poucos passos consecutivos. Outros critérios podem ser usados. No métodos das secantes, deve-se ter o cuidado de evitar divisões por zero quando $x_{n+1} - x_n$ muito pequeno em relação à resolução do sistema de numeração.

Table 3.4: Quadro comparativo.

Método	Convergência	Erro	Critério de parada
Bisseção	Linear $(p=1)$	$\epsilon_{n+1} = \frac{1}{2}\epsilon$	$\frac{b_n - a_n}{2} <$ erro
Iteração linear	Linear $(p=1)$	$\epsilon_{n+1} \approx \lvert\phi'(x^*)\rvert\varepsilon_n$	$\frac{\lvert\Delta_n\rvert}{1-\frac{\Delta_n}{\Delta_{n-1}}} <$ erro $\Delta_n < \Delta_{n-1}$
Newton	Quadrática $(p=2)$	$\epsilon_{n+1} \approx \frac{1}{2}\left\lvert\frac{f''(x^*)}{f'(x^*)}\right\rvert\varepsilon_n^2$	$\lvert\Delta_n\rvert <$ erro
Secante	$p = \frac{\sqrt{5}+1}{2} \approx 1{,}618$	$\varepsilon_{n+1} \approx \left\lvert\frac{f''(x^*)}{f'(x^*)}\right\rvert\varepsilon_n\varepsilon_{n-1} \approx M\varepsilon_n^{\phi}$	$\lvert\Delta_n\rvert <$ erro

Exercícios

E 3.6.1. Refaça as questões 3.4.3, 3.4.4, 3.4.5 e 3.4.6, usando o método das secantes.

E 3.6.2. Dê uma interpretação geométrica ao método das secantes. Qual a vantagem do método das secantes sobre o método de Newton?

E 3.6.3. Aplique o método das secantes para resolver a equação

$$e^{-x^2} = 2x \tag{3.200}$$

E 3.6.4. Refaça o Problema 3.2.8 usando o método de Newton e das secantes.

E 3.6.5. Seja uma função $f(x)$ dada duas vezes continuamente diferenciável. Faça uma análise assintótica para mostrar que as iterações do método das secantes satisfazem:

$$|x^{(n+1)} - x^*| \approx C|x^{(n)} - x^*||x^{(n-1)} - x^*|, \tag{3.201}$$

para aproximações iniciais $x^{(1)}$ e $x^{(2)}$ suficientemente próximas de x^*, onde $f(x^*) = 0$.

3.7 Exercícios finais

E 3.7.1. Calcule uma equação da reta tangente a curva $y = e^{-(x-1)^2}$ que passa pelo ponto $(3, 1/2)$.

E 3.7.2. Resolva numericamente a inequação:

$$e^{-x^2} < 2x \tag{3.219}$$

E 3.7.3. A equação

$$\cos(\pi x) = e^{-2x} \tag{3.220}$$

tem infinitas raízes. Usando métodos numéricos encontre as primeiras raízes dessa equação. Verifique a j-ésima raiz (z_j) pode ser aproximada por $j - 1/2$ para j grande. Use o método de Newton para encontrar uma aproximação melhor para z_j.

E 3.7.4. (Eletricidade) A corrente elétrica, I, em Ampéres em uma lâmpada em função da tensão elétrica, V, é dada por

$$I = \left(\frac{V}{150}\right)^{0.8} \tag{3.221}$$

Qual a potência da lâmpada quando ligada em série com uma resistência de valor R a uma fonte de 150V quando. (procure erro inferior a 1%)

a) $R = 0\Omega$

b) $R = 10\Omega$

c) $R = 50\Omega$

d) $R = 100\Omega$

E) $R = 500\Omega$

E 3.7.5. (Bioquímica) A concentração sanguínea de um medicamente é modelado pela seguinte expressão

$$c(t) = Ate^{-\lambda t} \tag{3.222}$$

onde $t > 0$ é o tempo em minutos decorrido desde a administração da droga. A é a quantidade administrada em mg/ml e λ é a constante de tempo em $\min^{-1}$. Responda:

a) Sendo $\lambda = 1/3$, em que instantes de tempo a concentração é metade do valor máximo. Calcule com precisão de segundos.

b) Sendo $\lambda = 1/3$ e $A = 100mg/ml$, durante quanto tempo a concentração permanece maior que $10mg/ml$.

E 3.7.6. Considere o seguinte modelo para crescimento populacional em um país:

$$P(t) = A + Be^{\lambda t}. \tag{3.223}$$

onde t é dado em anos. Use t em anos e $t = 0$ para 1960. Encontre os parâmetros A, B e λ com base nos anos de 1960, 1970 e 1991 conforme tabela:

Ano	população
1960	70992343
1970	94508583
1980	121150573
1991	146917459

Use esses parâmetros para calcular a população em 1980 e compare com o valor do censo. Dica: considere $\frac{P(31)-P(0)}{P(10)-P(0)}$ e reduza o sistema a uma equação apenas na variável λ.

E 3.7.7. (Fluidos) Uma boia esférica flutua na água. Sabendo que a boia tem 10ℓ de volume e 2Kg de massa. Calcule a altura da porção molhada da boia.

E 3.7.8. (Fluidos) Uma boia cilíndrica tem secção transversal circular de raio 10cm e comprimento 2m e pesa 10Kg. Sabendo que a boia flutua sobre água com o eixo do cilindro na posição horizontal, calcule a altura da parte molhada da boia.

E 3.7.9. Encontre com 6 casas decimais o ponto da curva $y = \ln x$ mais próximo da origem.

E 3.7.10. (Matemática financeira) Um computador é vendido pelo valor a vista de R$2.000,00 ou em 1+15 prestações de R$200,00. Calcule a taxa de juros associada à venda a prazo.

E 3.7.11. (Matemática financeira) O valor de R$110.000,00 é financiado conforme a seguinte programa de pagamentos:

Mês	pagamento
1	20.000,00
2	20.000,00
3	20.000,00
4	19.000,00
5	18.000,00
6	17.000,00
7	16.000,00

Calcule a taxa de juros envolvida. A data do empréstimo é o mês zero.

E 3.7.12. (Controle de sistemas) Depois de acionado um sistema de aquecedores, a temperatura em um forno evolui conforme a seguinte equação

$$T(t) = 500 - 800e^{-t} + 600e^{-t/3}. \tag{3.224}$$

onde T é a temperatura em Kelvin e t é tempo em horas.

a) Obtenha analiticamente o valor de $\lim_{t\to\infty} T(t)$.

b) Obtenha analiticamente o valor máximo de $T(t)$ e o instante de tempo quando o máximo acontece

c) Obtenha numericamente com precisão de minutos o tempo decorrido até que a temperatura passe pela primeira vez pelo valor de equilíbrio obtido no item a.

c) Obtenha numericamente com precisão de minutos a duração do período durante o qual a temperatura permanece pelo menos 20% superior ao valor de equilíbrio.

E 3.7.13. Encontre os pontos onde a elipse que satisfaz $\frac{x^2}{3} + y^2 = 1$ intersepta a parábola $y = x^2 - 2$.

E 3.7.14. (Otimização) Encontre a área do maior retângulo que é possível inscrever entre a curva $e^{-x^2}(1 + \cos(x))$ e o eixo $y = 0$.

E 3.7.15. (Otimização)Uma indústria consome energia elétrica de duas usinas fornecedoras. O custo de fornecimento em reais por hora como função da potência consumida em kW é dada pelas seguintes funções

$$C_1(x) = 500 + .27x + 4.1 \cdot 10^{-5}x^2 + 2.1 \cdot 10^{-7}x^3 + 4.2 \cdot 10^{-10}x^4 \quad (3.225)$$
$$C_2(x) = 1000 + .22x + 6.3 \cdot 10^{-5}x^2 + 8.5 \cdot 10^{-7}x^3 \quad (3.226)$$

Onde $C_1(x)$ e $C_2(x)$ são os custos de fornecimento das usinas 1 e 2, respectivamente. Calcule o custo mínimo da energia elétrica quando a potência total consumida é $1500kW$. Obs: Para um problema envolvendo mais de duas usinas, veja **??**.

E 3.7.16. (Termodinâmica) A pressão de saturação (em bar) de um dado hidrocarboneto pode ser modelada pela equação de Antoine:

$$\ln\left(P^{sat}\right) = A - \frac{B}{T + C} \quad (3.227)$$

onde T é a temperatura e A, B e C são constantes dadas conforme a seguir:

Hidrocarboneto	A	B	C
N-pentano	9.2131	2477.07	-39.94
N-heptano	9.2535	2911.32	-56.51

a) Calcule a temperatura de bolha de uma mistura de N-pentano e N-heptano à pressão de 1.2bar quando as frações molares dos gases são $z_1 = z_2 = 0.5$. Para tal utilize a seguinte equação:

$$P = \sum_i z_i P_i^{sat} \quad (3.228)$$

b) Calcule a temperatura de orvalho de uma mistura de N-pentano e N-heptano à pressão de 1.2bar quando as frações molares dos gases são $z_1 = z_2 = 0.5$. Para tal utilize a seguinte equação:

$$\frac{1}{P} = \sum_i \frac{z_i}{P_i^{sat}} \quad (3.229)$$

E 3.7.17. Encontre os três primeiros pontos de mínimo da função

$$f(x) = e^{-x/11} + x\cos(2x) \quad (3.230)$$

para $x > 0$ com erro inferior a 10^{-7}.

Chapter 4

Interpolação

Neste capítulo, discutimos os problemas de **interpolação**. Mais precisamente, dada uma sequência de n reais $x_1 < x_2 < \ldots < x_n$, um conjunto de pontos $\{(x_i, y_i) \in I \times \mathbb{R}\}_{i=1}^{n}$, onde $I = [x_1,x_n]$ e uma família de funções $\mathcal{F}_I = \{\varphi : I \to \mathbb{R}\}$, o problema de interpolação consiste em encontrar alguma função $f \in \mathcal{F}_I$ tal que

$$f(x_i) = y_i, \quad i = 1, 2, \ldots, n. \tag{4.1}$$

Chamamos uma tal f de **função interpoladora** dos pontos dados. Ou ainda, dizemos que f interpola os pontos dados.

Exemplo 4.0.1. Um dos problemas de interpolação mais simples é o de encontrar a equação da reta que passa por dois pontos dados. Por exemplo, sejam dados o conjunto de pontos $\{(1, 1), (2, 2)\}$ e a família de funções $\mathcal{F}_{[1,2]}$:

$$\mathcal{F}_{[1,2]} = \{f : [1,2] \to \mathbb{R} \ ; \ [1,2] \ni x \mapsto f(x) = a + bx; \ a,b \in \mathbb{R}\}. \tag{4.2}$$

Para que uma f na família seja a função interpoladora do conjunto de pontos dados, precisamos que

$$\begin{array}{l} a + bx_1 = y_1 \\ a + bx_2 = y_2 \end{array} \quad \text{isto é} \quad \begin{array}{l} a + b = 1 \\ a + 2b = 2 \end{array} \tag{4.3}$$

o que nos fornece $a = 0$ e $b = 1$. Então, a função interpoladora f é tal que $f(x) = x$ para um $x \in [1,2]$. Os pontos e a reta interpolada estão esboçados na Figura 4.1.

Um problema de interpolação cuja família de funções constitui-se de polinômios é chamado de problema de interpolação polinomial.

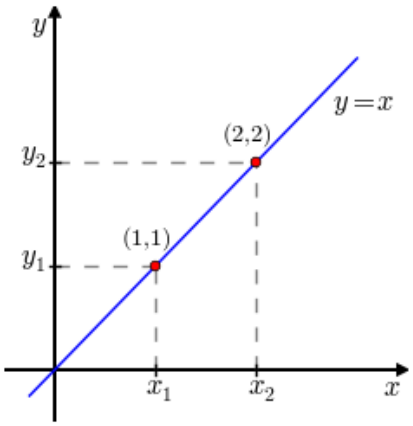


Figure 4.1: Exemplo de interpolação de dois pontos por uma reta, veja o Exemplo 4.0.1.

4.1 Interpolação polinomial

Interpolação polinomial é um caso particular do problema geral de interpolação, no qual a família de funções é constituída de polinômios. A escolha de polinômios como funções interpolantes é natural por diversos motivos, entre eles: se p é um polinômio de grau n, o valor $p(x)$ para um x real é calculado através de $n+1$ operações de multiplicação e $n+1$ operações de adição. Para tanto, pode-se usar o algoritmo de Horner[1]. Dado um polinômio p de grau n da forma

$$p(x) = \sum_{k=0}^{n} a_k x^k, \tag{4.4}$$

é possível reescrevê-lo como a sequência de operações dada por

$$a_0 + x\left(a_1 + x\left(a_2 + x\left(\ldots + x\left(a_{n-1} + xa_n\right)\ldots\right)\right)\right). \tag{4.5}$$

Também, derivadas e primitivas de polinômios são também polinômios cuja relação algébrica com o original é simples. Além disso, o teorema da aproximação de Weierstrass estabelece que qualquer função contínua definida em um intervalo fechado pode ser aproximada uniformemente por um polinômio tão bem quanto se queira.

Teorema 4.1.1 (Weierstrass)**.** *Seja f uma função contínua definida no intervalo fechado $[a,b]$ e seja δ um número positivo. Então existe um polinômio p, tal que para todo $x \in [a,b]$,*

$$|f(x) - p(x)| < \delta. \tag{4.6}$$

[1]William George Horner, 1786 - 1837, matemático britânico.

Observe que para o problema ser bem determinado, é necessário restringirmos o grau dos polinômios. Dado um conjunto de n pontos a serem interpolados $\{(x_i,y_i)\}_{i=1}^{n}$, $x_i \neq x_j$ para $i \neq j$, a família de polinômios $\mathcal{F} = \mathbb{P}_{n-1}$ deve ser escolhida, onde:

$$\mathbb{P}_{n-1} := \left\{ p : x \mapsto p(x) = \sum_{k=0}^{n-1} a_k x^k;\ \{a_0,a_1,\ldots,a_{n-1}\} \in \mathbb{R} \right\}, \tag{4.7}$$

isto é, a família dos polinômios reais de grau menor ou igual a $n-1$.

O Exemplo 4.0.1 discute um dos casos mais simples de interpolação polinomial, o qual consiste em interpolar uma reta por dois pontos. Neste caso, a família de funções consiste de polinômios de grau 1. Se buscarmos interpolar uma parábola pelos dois pontos dados, o problema fica subdeterminado, pois existem infinitas parábolas que passam por dois pontos dados. Além disso, se buscarmos interpolar uma reta por três pontos dados, o problema estaria sobredeterminado e poderia não ter solução se os pontos não fossem colineares. Veja o Exercício 4.1.1.

Assim, dado um conjunto com n pontos $\{(x_i,y_i)\}_{i=1}^{n}$, chamamos de **polinômio interpolador** o polinômio de grau menor ou igual a $n-1$ que os interpola.

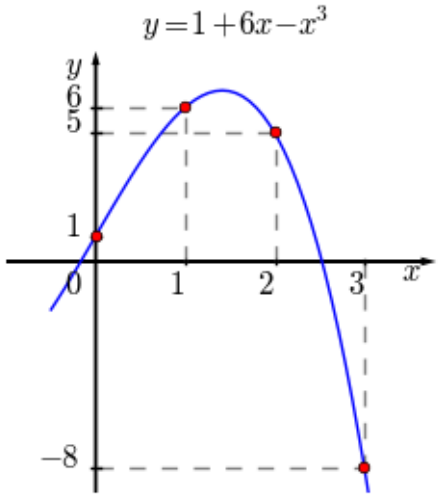


Figure 4.2: Polinômio interpolador do conjunto de pontos $\{(0,1),\ (1,6),\ (2,5),\ (3,-8)\}$. Veja o Exemplo 4.1.1.

Exemplo 4.1.1. Encontre o polinômio interpolador do conjunto de pontos $\{(0,1),\ (1,6),\ (2,5),\ (3,-8)\}$.

Solução. Como o conjunto consiste de 4 pontos, o polinômio interpolador deve ser da forma:

$$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3. \tag{4.8}$$

As condições de interpolação são $p(x_i) = y_i$, $i = 0, 1, 2, 3$, o que nos leva ao sistema linear:

$$\begin{array}{rcrcrcrcl} a_0 & & & & & & & = & 1 \\ a_0 & + & a_1 & + & a_2 & + & a_3 & = & 6 \\ a_0 & + & 2a_1 & + & 4a_2 & + & 8a_3 & = & 5 \\ a_0 & + & 3a_1 & + & 9a_2 & + & 27a_3 & = & -8 \end{array} \tag{4.9}$$

cuja solução é $a_0 = 1$, $a_1 = 6$, $a_2 = 0$ e $a_3 = -1$. Portanto, o polinômio interpolador é $p(x) = 1 + 6x - x^3$. Veja Figura 4.2.

No `Scilab`, podemos encontrar o polinômio interpolador e esboçar seu gráfico com os seguintes comandos:

```
-->xi = [0 1 2 3]';
-->yi = [1 6 5 -8]';
-->A = [xi.^0 xi.^1 xi.^2 xi.^3];
-->a = A\yi;
-->p = poly(a,'x','c')
  p  =
                3
     1 + 6x - x
-->xx = linspace(-0.5,3.25);
-->plot(xi,yi,'ro',xx,horner(p,xx),'b-');xgrid
```

◊

Teorema 4.1.2. *Seja* $\{(x_i,y_i)\}_{i=1}^n$ *um conjunto de* n *pares ordenados de números reais tais que* $x_i \neq x_j$ *se* $i \neq j$*, então existe um único polinômio* $p(x)$ *de grau* $n-1$ *ou inferior que passa por todos os pontos dados, isto é,* $p(x_i) = y_i, i = 1, \ldots, n$.

Proof. Observe que o problema de encontrar os coeficientes a_0, a_1,..., a_{n-1} do polinômio

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots a_{n-1}x^{n-1} = \sum_{k=0}^{n-1} a_kx^k \tag{4.10}$$

tal que $p(x_i) = y_i$ é equivalente a resolver o sistema linear com n equações e n incógnitas dado por

$$\begin{array}{c} a_0 + a_1x_1 + a_1x_1^2 + \cdots + a_{n-1}x_1^{n-1} = y_1, \\ a_0 + a_1x_2 + a_2x_2^2 + \cdots + a_{n-1}x_2^{n-1} = y_2, \\ \vdots \\ a_0 + a_1x_n + a_2x_n^2 + \cdots + a_{n-1}x_n^{n-1} = y_n. \end{array} \tag{4.11}$$

O qual pode ser escrito na forma matricial como

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ 1 & x_3 & x_3^2 & \cdots & x_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \tag{4.12}$$

A matriz envolvida é uma **matriz de Vandermonde**[2] de ordem n cujo determinante é dado pelo produtório duplo

$$\prod_{1 \leq i < j \leq n} (x_j - x_i) \tag{4.13}$$

É fácil ver que se as abscissas são diferentes dois a dois, então o determinante é não nulo. Disto decorre que a matriz envolvida é inversível e, portanto, o sistema possui uma solução que é única. □

Esta abordagem direta que usamos no Exemplo 4.1.1 e na demonstração do Teorema 4.1.2 se mostra ineficiente quando o número de pontos é grande e quando existe grande variação nas abscissas. Neste caso, a matriz de Vandermonde é mal condicionada (ver [**?**]), o que acarreta um aumento dos erros de arredondamento na solução do sistema.

Uma maneira de resolver este problema é escrever o polinômio em uma base que produza um sistema bem condicionado.

Exercícios resolvidos

Esta seção carece de exercícios resolvidos. Participe da sua escrita.

Veja como em:

https://github.com/livroscolaborativos/CalculoNumerico

ER 4.1.1. Mostre que:

a) Existem infinitas parábolas que interpolam dois pontos dados $\{(x_1, y_1), (x_2, y_2)\}$, com $x_1 \neq x_2$.

b) Não existe reta que interpola os pontos $\{(1, 1), (2, 2{,}1), (3, 3)\}$.

[2] Alexandre-Théophile Vandermonde, 1735 - 1796, matemático francês.

c) Não existe parábola de equação $y = a_0 + a_1x + a_2x^2$ que interpola dois pontos dados $\{(x_1, y_1), (x_1, y_2)\}$, com $y_1 \neq y_2$. Mas, existem infinitas parábolas de equação $x = a_0 + a_1y + a_2y^2$ que interpolam estes pontos.

Solução. a) Uma parábola de equação $y = a_1 + a_2x + a_3x^2$ que interpola os pontos deve satisfazer o sistema:

$$\begin{aligned} a_1 + a_2x_1 + a_3x_1^2 &= y_1 \\ a_1 + a_2x_2 + a_3x_2^2 &= y_2 \end{aligned} \tag{4.14}$$

Sem perda de generalidade, para cada $a_3 \in \mathbb{R}$ dado, temos:

$$\begin{aligned} a_1 + a_2x_1 &= y_1 - a_3x_1^2 \\ a_1 + a_2x_2 &= y_2 - a_3x_2^2, \end{aligned} \tag{4.15}$$

o qual tem solução única, pois $x_1 \neq x_2$. Ou seja, para cada $a_3 \in \mathbb{R}$ dado, existem $a_1, a_2 \in \mathbb{R}$ tais que a parábola de equação $y = a_1 + a_2x + a_3x^2$ interpola os pontos dados.

b) Certamente não existem retas de equação $x = a$ que interpolam os pontos dados. Consideremos então retas de equação $y = a_1 + a_2x$. Para uma tal reta interpolar os pontos dados é necessário que:

$$\begin{aligned} a_1 + a_2 &= 1 \\ a_1 + 2a_2 &= 2{,}1, \\ a_1 + 3a_2 &= 3 \end{aligned} \tag{4.16}$$

o qual é um sistema impossível.

c) Não existe uma parábola de equação $y = a_1 + a_2x + a_3x^2$ que interpole os pontos dados, pois tal equação determina uma função de x em y. Agora, para mostrar que existem infinitas parábolas de equação $x = a_1 + a_2y + a_3y^2$ que interpolam os pontos dados, basta seguir um raciocínio análogo ao do item a), trocando x por y e y por x.

$\Diamond$

Exercícios

Esta seção carece de exercícios. Participe da sua escrita.

Veja como em:

https://github.com/livroscolaborativos/CalculoNumerico

Prof. M.e Daniel Cassimiro

E 4.1.1. Encontre o polinômio interpolador para o conjunto de pontos $\{(-2,-47)$, $(0,-3)$, $(1,4)$, $(2,41)\}$. Então, faça um gráfico com os pontos e o polinômio interpolador encontrado.

E 4.1.2. Encontre o polinômio interpolador para o conjunto de pontos $\{(-1,1{,}25)$, $(0{,}5,0{,}5)$, $(1,1{,}25)$, $(1{,}25,1{,}8125)\}$.

4.2 Diferenças divididas de Newton

Dado um conjunto com n pontos $\{(x_i,y_i)\}_{i=1}^{n}$, o **método das diferenças divididas de Newton** consiste em construir o polinômio interpolador da forma

$$\begin{aligned} p(x) &= a_1 + a_2(x-x_1) + a_3(x-x_1)(x-x_2) + \cdots \\ &\quad + a_n(x-x_1)(x-x_2)\cdots(x-x_{n-1}). \end{aligned} \tag{4.17}$$

Como $p(x_i) = y_i$, $i = 1,2,\ldots,n$, os coeficientes a_i satisfazem o seguinte sistema triangular inferior:

$$\begin{aligned} a_1 &= y_1 \\ a_1 + a_2(x_2-x_1) &= y_2 \\ a_1 + a_2(x_3-x_1) + a_3(x_3-x_1)(x_3-x_2) &= y_3 \\ &\vdots \\ a_1 + a_2(x_n-x_1) + \cdots + a_n(x_n-x_1)\cdots(x_n-x_{n-1}) &= y_n \end{aligned} \tag{4.18}$$

Resolvendo de cima para baixo, obtemos

$$\begin{aligned} a_1 &= y_1 \\ a_2 &= \frac{y_2-a_1}{x_2-x_1} = \frac{y_2-y_1}{x_2-x_1} \\ a_3 &= \frac{y_3-a_2(x_3-x_1)-a_1}{(x_3-x_1)(x_3-x_2)} = \frac{\frac{y_3-y_2}{(x_3-x_2)} - \frac{y_2-y_1}{(x_2-x_1)}}{(x_3-x_1)} \\ &\cdots \end{aligned} \tag{4.19}$$

Note que os coeficientes são obtidos por diferenças das ordenadas divididas por diferenças das abscissas dos pontos dados. Para vermos isso mais claramente,

Table 4.1: Esquema de diferenças divididas para um conjunto com três pontos $\{(x_i, y_i)\}_{i=1}^{3}$.

j	x_j	$f[x_j]$	$f[x_{j-1},x_j]$	$f[x_{j-2},x_{j-1},x_j]$
1	x_1	$\boldsymbol{f[x_1]} = y_1$		
			$\boldsymbol{f[x_1,x_2]} = \dfrac{f[x_2]-f[x_1]}{x_2-x_1}$	
2	x_2	$f[x_2] = y_2$		$\boldsymbol{f[x_1,x_2,x_3]} = \dfrac{f[x_2,x_3]-f[x_1,x_2]}{x_3-x_1}$
			$f[x_2,x_3] = \dfrac{f[x_3]-f[x_2]}{x_3-x_2}$	
3	x_3	$f[x_3] = y_3$		

introduzimos a seguinte notação:

$$f[x_j] := y_j \tag{4.20}$$

$$f[x_j, x_{j+1}] := \frac{f[x_{j+1}] - f[x_j]}{x_{j+1} - x_j} \tag{4.21}$$

$$f[x_j, x_{j+1}, x_{j+2}] := \frac{f[x_{j+1}, x_{j+2}] - f[x_j, x_{j+1}]}{x_{j+2} - x_j} \tag{4.22}$$

$$\vdots \tag{4.23}$$

$$f[x_j, x_{j+1}, \ldots, x_{j+k}] := \frac{f[x_{j+1}, x_{j+2}, \ldots, x_{j+k}] - f[x_j, x_{j+1}, \ldots, x_{j+k-1}]}{x_{j+k} - x_j} \tag{4.24}$$

Chamamos $f[x_j]$ de diferença dividida de ordem zero (ou primeira diferença dividida), $f[x_i,x_j+1]$ de diferença dividida de ordem 1 (ou segunda diferença dividida) e assim por diante.

Uma inspeção cuidadosa dos coeficientes obtidos em (4.19) nos mostra que

$$a_k = f[x_1,x_2,\ldots,x_k] \tag{4.25}$$

Isto nos permite esquematizar o método conforme apresentado na Tabela 4.1.

Exemplo 4.2.1. Use o método de diferenças divididas para encontrar o polinômio que passe pelos pontos $(-1,3)$,$(0,1)$,$(1,3)$,$(3,43)$.

Solução. Usando o esquema apresentado na Tabela 4.1, obtemos

j	x_j	$f[x_j]$	$f[x_{j-1},x_j]$	$f[x_{j-2},x_{j-1},x_j]$	$f[x_{j-3},x_{j-2},x_{j-1},x_j]$
1	-1	**3**			
			$\frac{1-3}{0-(-1)} = \mathbf{-2}$		
2	0	1		$\frac{2-(-2)}{1-(-1)} = \mathbf{2}$	
			$\frac{3-1}{1-0} = 2$		$\frac{6-2}{3-(-1)} = \mathbf{1}$
3	1	3		$\frac{20-2}{3-0} = 6$	
			$\frac{43-3}{3-1} = 20$		
4	3	43			

Portanto, o polinômio interpolador do conjunto de pontos dados é

$$p(x) = 3 - 2(x+1) + 2(x+1)x + (x+1)x(x-1) \tag{4.26}$$

ou, equivalentemente, $p(x) = x^3 + 2x^2 - x + 1$.

$\Diamond$

4.3 Polinômios de Lagrange

Outra maneira clássica de resolver o problema da interpolação polinomial é através dos polinômios de Lagrange. Dado um conjunto de pontos $\{x_j\}_{j=1}^n$ distintos dois a dois, definimos os polinômios de Lagrange como os polinômios de grau $n-1$ que satisfazem

$$L_k(x_j) = \begin{cases} 1, & \text{se } k = j \\ 0, & \text{se } k \neq j \end{cases} \tag{4.27}$$

Assim, o polinômio $p(x)$ de grau $n-1$ que interpola os pontos dados, isto é, $p(x_j) = y_j, j = 1, \ldots, n$ é dado por

$$p(x) = y_1 L_1(x) + y_2 L_2(x) + \cdots + y_n L_n(x) = \sum_{k=1}^{n} y_k L_k(x). \tag{4.28}$$

Para construir os polinômios de Lagrange, podemos analisar a sua forma fatorada, ou seja:

$$L_k(x) = c_k \prod_{\substack{j=1 \\ j \neq k}}^{n} (x - x_j) \tag{4.29}$$

onde o coeficiente c_k é obtido da condição $L_k(x_k) = 1$:

$$L_k(x_k) = c_k \prod_{\substack{j=1 \\ j\neq k}}^{n} (x_k - x_j) \Longrightarrow c_k = \frac{1}{\prod_{\substack{j=1 \\ j\neq k}}^{n} (x_k - x_j)} \tag{4.30}$$

Portanto,

$$L_k(x) = \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{(x - x_j)}{(x_k - x_j)} \tag{4.31}$$

Observação 4.3.1. O problema de interpolação quando escrito usando como base os polinômios de Lagrange produz um sistema linear diagonal.

Exemplo 4.3.1. Encontre o polinômio da forma $p(x) = a_1 + a_2x + a_3x^2 + a_4x^3$ que passa pelos pontos $(0, 0)$, $(1, 1)$, $(2, 4)$, $(3, 9)$.

Solução. Escrevemos:

$$L_1(x) = \frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} = -\frac{1}{6}x^3 + x^2 - \frac{11}{6}x + 1 \tag{4.32}$$

$$L_2(x) = \frac{x(x-2)(x-3)}{1(1-2)(1-3)} = \frac{1}{2}x^3 - \frac{5}{2}x^2 + 3x \tag{4.33}$$

$$L_3(x) = \frac{x(x-1)(x-3)}{2(2-1)(2-3)} = -\frac{1}{2}x^3 + 2x^2 - \frac{3}{2}x \tag{4.34}$$

$$L_4(x) = \frac{x(x-1)(x-2)}{3(3-1)(3-2)} = \frac{1}{6}x^3 - \frac{1}{2}x^2 + \frac{1}{3}x \tag{4.35}$$

Assim, temos:

$$P(x) = 0 \cdot L_1(x) + 1 \cdot L_2(x) + 4 \cdot L_3(x) + 9 \cdot L_4(x) = x^2 \tag{4.36}$$

$\Diamond$

4.4 Aproximação de funções reais por polinômios interpoladores

Teorema 4.4.1. *Dados $n + 1$ pontos distintos, x_0, x_1, $\cdots$, x_n, dentro de um intervalo $[a,b]$ e uma função f com $n + 1$ derivadas contínuas nesse intervalo ($f \in C^{n+1}[a,b]$), então para cada x em $[a,b]$, existe um número $\xi(x)$ em (a,b) tal que*

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n), \tag{4.37}$$

onde $P(x)$ é o polinômio interpolador. Em especial, pode-se dizer que

$$|f(x) - P(x)| \leq \frac{M}{(n+1)!} |(x - x_0)(x - x_1) \cdots (x - x_n)|, \tag{4.38}$$

onde

$$M = \max_{x \in [a,b]} |f^{(n+1)}(\xi(x))| \tag{4.39}$$

Exemplo 4.4.1. Considere a função $f(x) = \cos(x)$ e o polinômio $P(x)$ de grau 2 tal que $P(0) = \cos(0) = 1$, $P(\frac{1}{2}) = \cos(\frac{1}{2})$ e $P(1) = \cos(1)$. Use a fórmula de Lagrange para encontrar $P(x)$. Encontre o erro máximo que se assume ao aproximar o valor de $\cos(x)$ pelo de $P(x)$ no intervalo [0,1]. Trace os gráficos de $f(x)$ e $P(x)$ no intervalo [0,1] no mesmo plano cartesiano e, depois, trace o gráfico da diferença $\cos(x) - P(x)$. Encontre o erro efetivo máximo $|\cos(x) - P(x)|$.

Solução. Usando polinômios de Lagrange, obtemos

$$\begin{aligned} P(x) &= 1\frac{(x - \frac{1}{2})(x - 1)}{(0 - \frac{1}{2})(0 - 1)} \\ &+ \cos\left(\frac{1}{2}\right) \frac{(x - 0)(x - 1)}{(\frac{1}{2} - 0)(\frac{1}{2} - 1)} \\ &+ \cos(1)\frac{(x - 0)(x - \frac{1}{2})}{(1 - 0)(1 - \frac{1}{2})} \qquad (4.40) \\ &\approx 1 - 0{,}0299720583066x - 0{,}4297256358252x^2 \qquad (4.41) \end{aligned}$$

No `Scilab`, podemos computar o polinômio interpolador da seguinte forma:

```
L1=poly([.5 1],'x');L1=L1/horner(L1,0)
L2=poly([0 1],'x');L2=L2/horner(L2,0.5)
L3=poly([0 .5],'x');L3=L3/horner(L3,1)
P=L1+cos(.5)*L2+cos(1)*L3
x=[0:.05:1]
plot(x,cos)
plot(x,horner(P,x),'red')
plot(x,horner(P,x)-cos(x))
```

Para estimar o erro máximo, precisamos estimar a derivada terceira de $f(x)$:

$$|f'''(x)| = |\operatorname{sen}(x)| \leq \operatorname{sen}(1) < 0{,}85 \tag{4.42}$$

e, assim,

$$\max_{x \in [0,1]} \left| x\left(x - \frac{1}{2}\right)(x - 1) \right|. \tag{4.43}$$

O polinômio de grau três $Q(x) = x\left(x - \frac{1}{2}\right)(x-1)$ tem um mínimo (negativo) em $x_1 = \frac{3+\sqrt{3}}{6}$ e um máximo (positivo) em $x_2 = \frac{3-\sqrt{3}}{6}$. Logo:

$$\max_{x\in[0,1]} \left| x\left(x - \frac{1}{2}\right)(x-1)\right| \leq \max\{|Q(x_1)|,\, |Q(x_2)|\} \approx 0{,}0481125. \tag{4.44}$$

Portanto:

$$|f(x) - P(x)| < \frac{0{,}85}{3!} 0{,}0481125 \approx 0{,}0068159 < 7 \cdot 10^{-3} \tag{4.45}$$

Para estimar o erro efetivo máximo, basta encontrar o máximo de $|P(x) - \cos(x)|$. O mínimo (negativo) de $P(x) - \cos(x)$ acontece em $x_1 = 4{,}29 \cdot 10^{-3}$ e o máximo (positivo) acontece em $x_2 = 3{,}29 \cdot 10^{-3}$. Portanto, o erro máximo efetivo é $4{,}29 \cdot 10^{-3}$. ♢

Exemplo 4.4.2. Considere o problema de aproximar o valor da integral $\int_0^1 f(x)dx$ pelo valor da integral do polinômio $P(x)$ que coincide com $f(x)$ nos pontos $x_0 = 0$, $x_1 = \frac{1}{2}$ e $x_2 = 1$. Use a fórmula de Lagrange para encontrar $P(x)$. Obtenha o valor de $\int_0^1 P(x)dx$ e encontre uma expressão para o erro de truncamento.

O polinômio interpolador de $f(x)$ é

$$\begin{aligned} P(x) &= f(0)\frac{(x-\frac{1}{2})(x-1)}{(0-\frac{1}{2})(0-1)} + f\left(\frac{1}{2}\right)\frac{(x-0)(x-1)}{(\frac{1}{2}-0)(\frac{1}{2}-1)} + f(1)\frac{(x-0)(x-\frac{1}{2})}{(1-0)(1-\frac{1}{2})} && (4.46)\\ &= f(0)(2x^2-3x+1) + f\left(\frac{1}{2}\right)(-4x^2+4x) + f(1)(2x^2-x) && (4.47)\end{aligned}$$

e a integral de $P(x)$ é:

$$\begin{aligned} \int_0^1 P(x)dx &= \left[f(0)\left(\frac{2}{3}x^3 - \frac{3}{2}x^2 + x\right)\right]_0^1 + \left[f\left(\frac{1}{2}\right)\left(-\frac{4}{3}x^3 + 2x^2\right)\right]_0^1 && (4.48)\\ &+ \left[f(1)\left(\frac{2}{3}x^3 - \frac{1}{2}x^2\right)\right]_0^1 && (4.49)\\ &= f(0)\left(\frac{2}{3} - \frac{3}{2} + 1\right) + f\left(\frac{1}{2}\right)\left(-\frac{4}{3} + 2\right) + f(1)\left(\frac{2}{3} - \frac{1}{2}\right) && (4.50)\\ &= \frac{1}{6}f(0) + \frac{2}{3}f\left(\frac{1}{2}\right) + \frac{1}{6}f(1) && (4.51)\end{aligned}$$

Para fazer a estimativa de erro usando o Teorema 4.4.1 e temos

$$\begin{aligned}
\left|\int_0^1 f(x)dx - \int_0^1 P(x)dx\right| &= \left|\int_0^1 f(x) - P(x)dx\right| && (4.52)\\
&\leq \int_0^1 |f(x) - P(x)|dx && (4.53)\\
&\leq \frac{M}{6}\int_0^1 \left|x\left(x-\frac{1}{2}\right)(x-1)\right|dx && (4.54)\\
&= \frac{M}{6}\left[\int_0^{1/2} x\left(x-\frac{1}{2}\right)(x-1)dx\right. && (4.55)\\
&- \left.\int_{1/2}^1 x\left(x-\frac{1}{2}\right)(x-1)dx\right] && (4.56)\\
&= \frac{M}{6}\left[\frac{1}{64} - \left(-\frac{1}{64}\right)\right] = \frac{M}{192}. && (4.57)
\end{aligned}$$

Lembramos que $M = \max_{x\in[0,1]} |f'''(x)|$.

Observação 4.4.1. Existem estimativas melhores para o erro de truncamento para este esquema de integração numérica. Veremos com mais detalhes tais esquemas na teoria de integração numérica.

Exemplo 4.4.3. Use o resultado do exemplo anterior para aproximar o valor das seguintes integrais:

a) $\int_0^1 \ln(x+1)dx$

b) $\int_0^1 e^{-x^2}dx$

Solução. Usando a fórmula obtida, temos que

$$\int_0^1 \ln(x+1)dx \approx 0{,}39 \pm \frac{1}{96} \quad (4.58)$$

$$\int_0^1 e^{-x^2}dx \approx 0{,}75 \pm \frac{3{,}87}{192} \quad (4.59)$$

◊

Exercícios

E 4.4.1. Use as mesmas técnicas usadas o resultado do Exemplo 4.4.2 para obter uma aproximação do valor de:

$$\int_0^1 f(x)dx \quad (4.60)$$

através do polinômio interpolador que coincide com $f(x)$ nos pontos $x = 0$ e $x = 1$.

4.5 Interpolação linear segmentada

Considere o conjunto $(x_i,y_i)_{j=1}^{n}$ de n pontos. Assumiremos que $x_{i+1} > x_i$, ou seja, as abscissas são distintas e estão em ordem crescente. A função linear que interpola os pontos x_i e x_{i+1} no intervalo i é dada por

$$P_i(x) = y_i \frac{(x_{i+1} - x)}{(x_{i+1} - x_i)} + y_{i+1} \frac{(x - x_i)}{(x_{i+1} - x_i)} \tag{4.61}$$

O resultado da interpolação linear segmentada é a seguinte função contínua definida por partes no intervalo $[x_1,x_n]$:

$$f(x) = P_i(x), \qquad x \in [x_i,x_{i+1}] \tag{4.62}$$

Exemplo 4.5.1. Construa uma função linear por partes que interpola os pontos (0,0), (1,4), (2,3), (3,0), (4,2), (5,0).

A função procurada pode ser construída da seguinte forma:

$$f(x) = \begin{cases} 0\frac{x-1}{0-1} + 1\frac{x-0}{1-0}, & 0 \leq x < 1 \\ 4\frac{x-2}{1-2} + 3\frac{x-1}{2-1}, & 1 \leq x < 2 \\ 3\frac{x-3}{2-3} + 0\frac{x-2}{3-2}, & 2 \leq x < 3 \\ 0\frac{x-4}{3-4} + 2\frac{x-3}{4-3}, & 3 \leq x < 4 \\ 2\frac{x-5}{4-5} + 0\frac{x-4}{5-4}, & 4 \leq x \leq 5 \end{cases} \tag{4.63}$$

Simplificando, obtemos:

$$f(x) = \begin{cases} x, & 0 \leq x < 1 \\ -x + 5, & 1 \leq x < 2 \\ -3x + 9, & 2 \leq x < 3 \\ 2x - 6, & 3 \leq x < 4 \\ -2x + 10, & 4 \leq x \leq 5 \end{cases} \tag{4.64}$$

A Figura 4.3 é um esboço da função $f(x)$ obtida.

Ela foi gerada no `Scilab` usando os comandos:

```
//pontos fornecidos
xi = [0;1;2;3;4;5]
yi = [0;4;3;0;2;0]
```

```
//numero de pontos
n = 6
//funcao interpoladora
function [y] = f(x)
  for i=1:n-2
    if ((x>=xi(i)) & (x<xi(i+1))) then
      y = yi(i)*(x-xi(i+1))/(xi(i) - xi(i+1)) ...
          + yi(i+1)*(x-xi(i))/(xi(i+1) - xi(i));
    end
  end

  if ((x>=xi(n-1)) & (x<=xi(n))) then
    y = yi(n-1)*(x-xi(n))/(xi(n-1) - xi(n)) ...
        + yi(n)*(x-xi(n-1))/(xi(n) - xi(n-1));
  end
endfunction
//graficando
xx = linspace(xi(1),xi(n),500)';
clear yy
for i=1:max(size(xx))
  yy(i) = f(xx(i))
end
plot(xi,yi,'r.',xx,yy,'b-')
```

4.6 Interpolação cúbica segmentada - spline

A ideia empregada na interpolação linear segmentada pode ser estendida através da utilização de polinômios de grau superior. A escolha de polinômios de grau superior implica uma maior liberdade (há um número maior de coeficientes) na construção da interpolação. Parte dessa liberdade pode ser utilizada na exigência de suavidade para a interpolação.

Definição 4.6.1 (spline de ordem m)**.** *Dado um conjunto de n pontos $\mathcal{I} = \{(x_j,y_j)\}_{j=1}^{n}$ tais que $x_{j+1} > x_j$, ou seja, as abscissas são distintas e estão em ordem crescente; um spline de ordem m que interpola estes pontos é uma função s com as seguintes propriedades:*

i) Em cada intervalo $[x_j,x_{j+1})$, $j = 1,2,\ldots n-2$ e no segmento $[x_{n-1},x_n]$ s é um polinômio de grau menor ou igual a m;

ii) Em algum dos intervalos s é um polinômio de grau m;

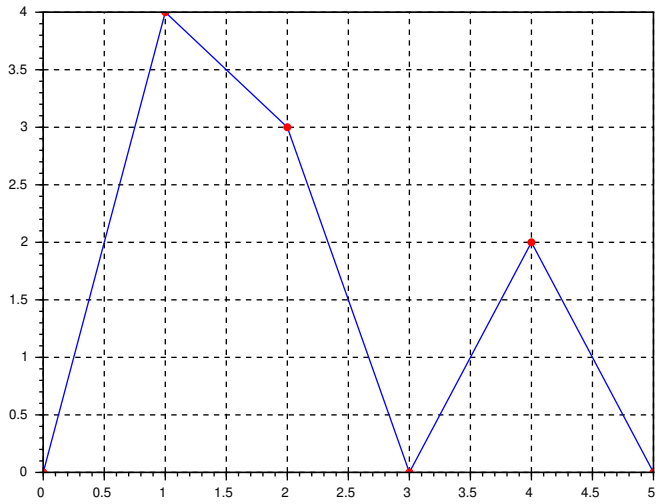


Figure 4.3: Interpolação linear segmentada.

iii) *Em cada* $x_j \in \mathcal{I}$, $s(x_j) = y_j$, *isto é, o spline interpola os pontos dados;*

iv) s *é uma função de classe* $\mathcal{C}^{m-1}$, *isto é, é função* $m-1$ *vezes continuamente diferenciável.*

São $n-1$ intervalos e em cada um deles há $m+1$ coeficientes a se determinar. As condições *iii* e *iv* impostas pela definição correspondem respectivamente a n e $m(n-2)$ equações. Estas últimas, se devem à exigência de continuidade nos pontos internos, ou seja, os pontos de $\mathcal{I}$ com índices $j = 2,3,\ldots,n-1$. Portanto, há $m-1$ coeficientes a mais do que o número de equações e, à exceção do caso $m = 1$ (interpolação linear segmentada), o problema é subdeterminado. Ou seja, uma vez fixada a ordem $m > 1$, existem infinitos splines de ordem m que interpolam os pontos do conjunto $\mathcal{I}$.

O caso $m = 3$, denominado spline cúbico, é de grande interesse pois reproduz o comportamento físico de réguas delgadas com estrutura elástica homogênea e perfil uniforme sujeitas aos vínculos representados pelos pontos do conjunto $\mathcal{I}$. A equação diferencial que rege o comportamento do perfil dessas réguas é um caso particular do equação da viga de Euler-Bernoulli. Neste caso, a equação tem a forma

$$\frac{d^4y}{dx^4} = 0, \tag{4.65}$$

cuja solução geral é um polinômio de grau 3.

Vamos supor que um spline cúbico que interpola o conjunto de pontos $\mathcal{I}$ é conhecido. Como esse spline é uma função de classe $\mathcal{C}^2$, as suas derivadas nos pontos do conjunto $\mathcal{I}$ são conhecidas também. Seja y'_j, o valor dessa derivada em $x = x_j$. Agora, vamos considerar dois pares de pontos sucessivos de $\mathcal{I}$, (x_j,y_j) e (x_{j+1},y_{j+1}). A forma do spline cúbico no intervalo $[x_j,x_{j+1})$ pode ser identificada com a solução da equação diferencial (4.65) no intervalo (x_j,x_{j+1}) sujeita às condições de contorno

$$y(x_j) = y_j, \quad y'(x_j) = y'_j, \quad y(x_{j+1}) = y_{j+1} \quad \text{e} \quad y'(x_{j+1}) = y'_{j+1}. \tag{4.66}$$

A solução desse problema de contorno é escrita de modo conveniente como

$$s_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \tag{4.67}$$

onde as constantes a_j, b_j, c_j e d_j se relacionam às do problema de contorno. As duas primeiras seguem imediatamente das condições de contorno em x_j:

$$a_j = y_j \quad \text{e} \quad b_j = y'_j. \tag{4.68}$$

As duas últimas são obtidas pela solução do sistema de equações formado pelas condições de contorno em x_{j+1}:

$$c_j = 3\frac{y_{j+1} - y_j}{(x_{j+1} - x_j)^2} - \frac{y'_{j+1} + 2y'_j}{x_{j+1} - x_j} \quad \text{e} \quad d_j = -2\frac{y_{j+1} - y_j}{(x_{j+1} - x_j)^3} + \frac{y'_{j+1} + y'_j}{(x_{j+1} - x_j)^2} \tag{4.69}$$

Esta relação entre o conjunto de valores para a derivada de um spline cúbico $\{y'_j\}_{j=1}^n$ nos pontos de interpolação $\mathcal{I}$ e os coeficientes dos polinômios em cada intervalo de interpolação pode ser resumida na seguinte proposição:

Proposição 4.6.1. *Seja s um spline cúbico que interpola o conjunto de pontos $\mathcal{I} = \{(x_j,y_j)\}_{j=1}^n \subset \mathbb{R}^2$ tais que $x_{j+1} > x_j$. Se $\{y'_j\}_{j=1}^n$ é o conjunto dos valores da derivada de s em x_j, então em cada intervalo $[x_j,x_{j+1})$ (fechado também à direita quando $j = n - 1$) o spline é igual a s_j:*

$$s_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \tag{4.70}$$

onde

$$\begin{aligned} a_j &= y_j, \quad c_j = 3\frac{y_{j+1} - y_j}{h_j^2} - \frac{y'_{j+1} + 2y'_j}{h_j}, \\ b_j &= y'_j, \quad d_j = -2\frac{y_{j+1} - y_j}{h_j^3} + \frac{y'_{j+1} + y'_j}{h_j^2} \end{aligned} \tag{4.71}$$

e

$$h_j = x_{j+1} - x_j, \quad j = 1,2,\ldots,n-1 \tag{4.72}$$

é a distância entre as abscissas de dois pontos de interpolação consecutivos.

De acordo com a proposição anterior, toda informação sobre um spline cúbico é armazenada no conjunto $\{(x_j,y_j,y'_j)\}_{j=1}^n$. Por construção, uma função s definida a partir de (4.70), (4.71) e (4.72) com um conjunto $\{(x_j,y_j,y'_j)\}_{j=1}^n \subset \mathbb{R}^3$, onde $x_{j+1} > x_j$ é de classe $\mathcal{C}^1$ mas não necessariamente um spline cúbico. Para ser um spline cúbico, os valores do conjunto $\{y'_j\}_{j=1}^n$ devem garantir a continuidade da derivada segunda de s em todo intervalo (x_1,x_n). Ou seja, devemos ter

$$\lim_{x \nearrow x_{j+1}} s''_j(x) = s''_{j+1}(x_{j+1}) \tag{4.73}$$

em todos os pontos internos $j = 1,2,\ldots,n-2$. Em termos dos coeficientes dos polinômios cúbicos (4.70), a equação anterior assume a forma

$$2c_j + 6d_j h_j = 2c_{j+1}, \quad j = 1,2,\ldots,n-2. \tag{4.74}$$

Esta última equação e (4.71) permitem construir um sistema de equações lineares para as variáveis y'_j:

Proposição 4.6.2. *Dado o conjunto de pontos $\mathcal{I} = \{(x_j,y_j)\}_{j=1}^n \subset \mathbb{R}^2$ tais que $x_{j+1} > x_j$, as derivadas de um spline cúbico que interpola os pontos $\mathcal{I}$, y'_j, $j = 1,2,\ldots,n$ satisfazem o sistema de equações algébricas lineares*

$$h_j y'_{j-1} + 2(h_{j-1}+h_j)y'_j + h_{j-1}y'_{j+1} = 3\left(h_j \frac{y_j - y_{j-1}}{h_{j-1}} + h_{j-1}\frac{y_{j+1}-y_j}{h_j}\right), \tag{4.75}$$

onde $j = 2,3,\ldots,n-1$ e $h_j = x_{j+1} - x_j$.

O sistema de equações (4.75) é subdeterminado. São n variáveis e $n-2$ equações. A inclusão de duas equações adicionais linearmente independentes das $n-2$ equações (4.75) possibilita a existência de uma única solução. Tipicamente essas equações adicionais envolvem o comportamento do spline na fronteira ou na sua vizinhança. A seguir, veremos quatro escolhas mais conhecidas.

4.6.1 Spline natural

Uma forma de definir as duas equações adicionais para completar o sistema (4.75) é impor condições de fronteira livres (ou naturais), ou seja,

$$s''(x_1) = s''(x_n) = 0. \tag{4.76}$$

De acordo com (4.70) essas equações implicam respectivamente

$$c_1 = 0 \quad \text{e} \quad 2c_{n-1} + 6d_{n-1}h_{n-1} = 0, \tag{4.77}$$

ou seja,

$$\begin{cases} 2y_1' + y_2' = 3\dfrac{y_2 - y_1}{h_1} \\ \\ y_{n-1}' + 2y_n' = 3\dfrac{y_n - y_{n-1}}{h_{n-1}} \end{cases}. \tag{4.78}$$

Essas duas equações em conjunto com as equações (4.75) formam um sistema de n equações algébricas lineares $Ay' = z$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 & \cdots & 0 & 0 \\ h_2 & 2(h_1+h_2) & h_1 & 0 & \cdots & 0 & 0 \\ 0 & h_3 & 2(h_2+h_3) & h_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-1} & 2(h_{n-1}+h_{n-2}) & h_{n-2} \\ 0 & 0 & 0 & \cdots & 0 & 1 & 2 \end{bmatrix}, \tag{4.79}$$

$$y' = \begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix} \quad \text{e} \quad z = 3\begin{bmatrix} \frac{y_2-y_1}{h_1} \\ h_2\frac{y_2-y_1}{h_1} + h_1\frac{y_3-y_2}{h_2} \\ h_3\frac{y_3-y_2}{h_2} + h_2\frac{y_4-y_3}{h_3} \\ \vdots \\ h_{n-1}\frac{y_{n-1}-y_{n-2}}{h_{n-2}} + h_{n-2}\frac{y_n-y_{n-1}}{h_{n-1}} \\ \frac{y_n-y_{n-1}}{h_{n-1}} \end{bmatrix}. \tag{4.80}$$

Observe que a matriz A é diagonal dominante estrita e, portanto, o sistema $Ay' = z$ possui solução única. Calculado y', os valores dos a_j, b_j, c_j e d_j são obtidos diretamente pelas expressões (4.71).

Exemplo 4.6.1. Construa um spline cúbico natural que passe pelos pontos (2, 4,5), (5, − 1,9), (9, 0,5) e (12, − 0,5).

Solução. O spline desejado é uma função definida por partes da forma:

$$s(x) = \begin{cases} a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3, & 2 \le x < 5 \\ a_2 + b_2(x-5) + c_2(x-5)^2 + d_2(x-5)^3, & 5 \le x < 9 \\ a_3 + b_3(x-9) + c_3(x-9)^2 + d_3(x-9)^3, & 9 \le x \le 12 \end{cases}. \tag{4.81}$$

As variáveis y'_1, y'_2, y'_3 e y'_4 resolvem o sistema $Ay' = z$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 4 & 2(4+3) & 3 & 0 \\ 0 & 3 & 2(3+4) & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 4 & 14 & 3 & 0 \\ 0 & 3 & 14 & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix}, \tag{4.82}$$

$$y = \begin{bmatrix} y'_1 \\ y'_2 \\ y'_3 \\ y'_4 \end{bmatrix} \quad \text{e} \quad z = 3 \begin{bmatrix} \frac{1}{3}(-1{,}9 - 4{,}5) \\ \frac{4}{3}(-1{,}9 - 4{,}5) + \frac{3}{4}(0{,}5 - (-1{,}9)) \\ \frac{3}{4}(0{,}5 - (-1{,}9)) + \frac{4}{3}(-0.5 - (0{,}5)) \\ \frac{1}{3}(-0{,}5 - (0.5)) \end{bmatrix} = \begin{bmatrix} -6{,}4 \\ -20{,}2 \\ 1{,}4 \\ -1 \end{bmatrix}. \tag{4.83}$$

A solução é $y'_1 = -2{,}8\bar{3}$, $y'_2 = -0{,}7\bar{3}$, $y'_3 = 0{,}4\bar{6}$ e $y'_4 = -0{,}7\bar{3}$. Calculamos os coeficientes usando as expressões (4.71):

$$\begin{array}{lllllll} a_1 & = & y_1 = 4{,}5, & & b_1 & = & y'_1 = -2{,}8\bar{3}, \\ a_2 & = & y_2 = -1{,}9, & & b_2 & = & y'_2 = -0{,}7\bar{3}, \\ a_3 & = & y_3 = 0{,}5. & & b_3 & = & y'_3 = 0{,}4\bar{6}, \end{array} \tag{4.84}$$

$$\begin{array}{lllllll} c_1 & = & 0, & & d_1 & = & 0{,}0\bar{7}, \\ c_2 & = & 0{,}7, & & d_2 & = & -0{,}091\bar{6}, \\ c_3 & = & -0{,}4, & & d_3 & = & 0{,}0\bar{4}. \end{array}$$

Portanto:

$$S(x) = \begin{cases} 4{,}5 - 2{,}8\bar{3}(x-2) + 0{,}0\bar{7}(x-2)^3 & , 2 \le x < 5 \\ -1{,}9 - 0{,}7\bar{3}(x-5) + 0{,}7(x-5)^2 - 0{,}091\bar{6}(x-5)^3 & , 5 \le x < 9 \\ 0{,}5 + 0{,}4\bar{6}(x-9) - 0{,}4(x-9)^2 + 0{,}0\bar{4}(x-9)^3 & , 9 \le x \le 12 \end{cases}. \tag{4.85}$$

No `Scilab`, podemos utilizar:

```
xi = [2;5;9;12]
yi = [4.5;-1.9;0.5;-0.5]
hi = xi(2:4)-xi(1:3)
A = [2 1 0 0;hi(2) 2*(hi(1)+hi(2)) hi(1) 0; ...
```

```
   0 hi(3) 2*(hi(2)+hi(3)) hi(2);0 0 1 2 ]
z = 3*[(yi(2)-yi(1))/hi(1); ...
   hi(2)/hi(1)*(yi(2)-yi(1))+hi(1)/hi(2)*(yi(3)-yi(2));...
   hi(3)/hi(2)*(yi(3)-yi(2))+hi(2)/hi(3)*(yi(4)-yi(3));...
   (yi(4)-yi(3))/hi(3)]
dyi = A\z
a=yi(1:3)
b=dyi(1:3)
c(1)=0
c(2:3)=3*(yi(3:4)-yi(2:3))./hi(2:3).^2 ...
         - (dyi(3:4)+2*dyi(2:3))./hi(2:3)
d=-2*(yi(2:4)-yi(1:3))./hi.^3 + (dyi(2:4)+dyi(1:3))./hi.^2
for i=1:3
    P(i) = poly([a(i) b(i) c(i) d(i)],'x','coeff')
    z = [xi(i):.01:xi(i+1)]
    plot(z,horner(P(i),z-xi(i)))
end
```

O mesmo resultado é obtido através das instruções `splin` e `interp` do `Scilab`:

```
xi = [2;5;9;12]
yi = [4.5;-1.9;0.5;-0.5]
dyi=splin(xi,yi,'natural')
z=linspace(xi(1),xi($))
plot(z,interp(z,xi,yi,dyi))
```

◊

4.6.2 Spline fixado

O spline fixado s é obtido pela escolha dos valores das derivadas nas extremidades do intervalo de interpolação. Isto diminui o número de variáveis para $n-2$ pois y_1' e y_n' deixam de ser incógnitas.

As equações (4.75) formam um sistema de $n-2$ equações $Ay' = z$, onde

$$A = \begin{bmatrix} 2(h_1+h_2) & h_1 & 0 & 0 & \cdots & 0 & 0 \\ h_3 & 2(h_2+h_3) & h_2 & 0 & \cdots & 0 & 0 \\ 0 & h_4 & 2(h_3+h_4) & h_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-2} & 2(h_{n-3}+h_{n-2}) & h_{n-3} \\ 0 & 0 & 0 & \cdots & 0 & h_{n-1} & 2(h_{n-2}+h_{n-1}) \end{bmatrix}, \tag{4.86}$$

$$y' = \begin{bmatrix} y'_2 \\ y'_3 \\ \vdots \\ y'_{n-1} \end{bmatrix} \quad \text{e} \quad z = 3 \begin{bmatrix} h_2 \frac{y_2 - y_1}{h_1} + h_1 \frac{y_3 - y_2}{h_2} - h_2 y'_1 \\ h_3 \frac{y_3 - y_2}{h_2} + h_2 \frac{y_4 - y_3}{h_3} \\ \vdots \\ h_{n-2} \frac{y_{n-2} - y_{n-3}}{h_{n-3}} + h_{n-3} \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\ h_{n-1} \frac{y_{n-1} - y_{n-2}}{h_{n-2}} + h_{n-2} \frac{y_n - y_{n-1}}{h_{n-1}} - h_{n-2} y'_n \end{bmatrix}. \tag{4.87}$$

Observe que a matriz A é diagonal dominante estrita e, portanto, o sistema $Ay' = z$ possui solução única.

4.6.3 Spline *not-a-knot*

O spline *not-a-knot* é definido com um spline cúbico que satisfaz as equações adicionais

$$\lim_{x \nearrow x_2} s'''_1(x) = s'''_2(x_2) \quad \text{e} \quad \lim_{x \nearrow x_{n-1}} s'''_{n-2}(x) = s'''_{n-1}(x_{n-1}). \tag{4.88}$$

Em termos dos coeficientes (4.70), as equações anteriores correspondem a

$$d_1 = d_2 \quad \text{e} \quad d_{n-2} = d_{n-1}, \tag{4.89}$$

ou seja,

$$\begin{cases} h_2^2 y'_1 + (h_2^2 - h_1^2) y'_2 - h_1^2 y'_3 = 2 \left(h_2^2 \dfrac{y_2 - y_1}{h_1} - h_1^2 \dfrac{y_3 - y_2}{h_2} \right) \\ \\ h_{n-1}^2 y'_{n-2} + (h_{n-1}^2 - h_{n-2}^2) y'_{n-1} - h_{n-2}^2 y'_n = 2 \left(h_{n-1}^2 \dfrac{y_{n-1} - y_{n-2}}{h_{n-2}} - h_{n-2}^2 \dfrac{y_n - y_{n-1}}{h_{n-1}} \right) \end{cases}. \tag{4.90}$$

Essas duas equações agregadas às equações (4.75) formam um sistema de n equações $Ay' = z$, onde

$$A = \begin{bmatrix} h_2^2 & h_2^2 - h_1^2 & -h_1^2 & 0 & \cdots & 0 & 0 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 & \cdots & 0 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-1} & 2(h_{n-2} + h_{n-1}) & h_{n-2} \\ 0 & 0 & 0 & \cdots & h_{n-1}^2 & h_{n-1}^2 - h_{n-2}^2 & -h_{n-2}^2 \end{bmatrix}, \tag{4.91}$$

$$y' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{bmatrix} \quad \text{e} \quad z = \begin{bmatrix} 2\left(h_2^2 \frac{y_2 - y_1}{h_1} - h_1^2 \frac{y_3 - y_2}{h_2}\right) \\ 3\left(h_2 \frac{y_2 - y_1}{h_1} + h_1 \frac{y_3 - y_2}{h_2}\right) \\ \vdots \\ 3\left(h_{n-1} \frac{y_{n-1} - y_{n-2}}{h_{n-2}} + h_{n-2} \frac{y_n - y_{n-1}}{h_{n-1}}\right) \\ 2\left(h_{n-1}^2 \frac{y_{n-1} - y_{n-2}}{h_{n-2}} - h_{n-2}^2 \frac{y_n - y_{n-1}}{h_{n-1}}\right) \end{bmatrix}. \tag{4.92}$$

Se reduzirmos esse sistema pela eliminação das incógnitas y'_1 e y'_n, o sistema resultante possui uma matriz de coeficientes diagonal dominante estrita, portanto, a solução é única.

O termo *not-a-knot* (não nó) relaciona-se à nomenclatura dos splines. O termo *nó* é utilizado para os pontos interpolados. Neles, a derivada terceira da função spline é descontínua, portanto, quando impomos a continuidade dessa derivada em x_2 e x_{n-1} é como se esses pontos deixassem de ser nós.

4.6.4 Spline periódico

Se o conjunto de n pontos da interpolação $\mathcal{I}$ for tal que $y_1 = y_n$, então é possível construir o spline periódico, definido com um spline cúbico que satisfaz as seguintes condições de periodicidade

$$s'_1(x_1) = s'_{n-1}(x_n) \quad \text{e} \quad s''_1(x_1) = s''_{n-1}(x_n). \tag{4.93}$$

Em termos dos coeficientes (4.70)

$$b_1 = b_{n-1} \quad \text{e} \quad 2c_1 = 2c_{n-1} + 6d_{n-1}h_{n-1}, \tag{4.94}$$

ou seja,

$$\begin{cases} y'_1 - y'_n = 0 \\ \\ 2h_{n-1}y'_1 + h_{n-1}y'_2 + h_1 y'_{n-1} + 2h_1 y'_n = 3\left(h_{n-1}\dfrac{y_2 - y_1}{h_1} + h_1 \dfrac{y_n - y_{n-1}}{h_{n-1}}\right) \end{cases}. \tag{4.95}$$

Essas duas equações agregadas às equações (4.75) formam um sistema de n equações $Ay' = z$, onde

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & -1 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 & \cdots & 0 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-1} & 2(h_{n-2} + h_{n-1}) & h_{n-2} \\ 2h_{n-1} & h_{n-1} & 0 & \cdots & 0 & h_1 & 2h_1 \end{bmatrix}, \tag{4.96}$$

$$y' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{bmatrix} \quad \text{e} \quad z = 3 \begin{bmatrix} 0 \\ h_2 \frac{y_2 - y_1}{h_1} + h_1 \frac{y_3 - y_2}{h_2} \\ \vdots \\ h_{n-1} \frac{y_{n-1} - y_{n-2}}{h_{n-2}} + h_{n-2} \frac{y_n - y_{n-1}}{h_{n-1}} \\ h_{n-1} \frac{y_2 - y_1}{h_1} + h_1 \frac{y_n - y_{n-1}}{h_{n-1}} \end{bmatrix}. \tag{4.97}$$

Neste caso também, se reduzirmos esse sistema pela eliminação das incógnitas y'_1 e y'_n, o sistema resultante possui uma matriz de coeficientes diagonal dominante estrita, portanto, a solução é única.

Appendix A

Rápida introdução à linguagem Julia

Neste apêndice, abordaremos a linguagem computacional `Julia` que nos auxiliará no processo de obtenção de resultados de operações e computações utilizando o computador.

A.1 Sobre a linguagem Julia

A linguagem de programação Julia é uma linguagem moderna e poderosa que foi criada para atender aos requisitos da computação numérica de alto desempenho e científica. Seus criadores quiseram reunir na linguagem as principais (melhores) características de outras linguagens:

- Ruby;
- matlab;
- R;
- Julia;
- C.

Julia é uma linguagem de programação de alto nível, interpretada e multi-paradigma.

Para mais informações, consulte:

- Página oficial da linguagem Julia: `https://julialang.org/`
- Manual Julia: `https://docs.julialang.org/en/v1/manual/getting-started/`

- GitHub do Professor Daniel: https://github.com/Daniel-C-Fernandes/Numerico/blob/main/Mini-curso-Julia.ipynb
- Julia for Data science: https://juliadatascience.io/pt/
- Julia Academy: https://juliaacademy.com/
- GitHub Julia: https://github.com/JuliaLang/julia
- Introdução: https://julia.vento.eng.br/
- Curso Julia USP: https://edisciplinas.usp.br/pluginfile.php/7879038/mod_resource/content/4/Tutorial%20para%20a%20Linguagem%20Julia.pdf
- Tutoriais Julia: https://julialang.org/learning/tutorials/

Para saber mais: https://www.youtube.com/watch?v=Xzdg9czOxD8&t=29s

A.1.1 Características Principais

- **Tipagem Dinâmica:** Suas variáveis podem receber qualquer tipo de dado, e sua sintaxe se aproxima mais da linguagem humana do que da linguagem de máquina.
- **Multiparadigma:** Suporta diversos paradigmas de programação, como orientação a objetos e programação funcional.
- **Alto Nível:** Possui uma sintaxe expressiva e amigável.
- **Gratuita e Open Source:** Julia é distribuída sob a licença MIT.
- **Suporte a Unicode e UTF-8:** Permite o uso de símbolos matemáticos durante a escrita de programas. Gerenciador de Pacotes Prático: Facilita a instalação e atualização de pacotes.

Aplicações:

- **Data Science:** Julia é usada para análise de dados e descoberta de conhecimento a partir de grandes conjuntos de dados. Machine Learning: Possui pacotes poderosos para criação de modelos de aprendizado de máquina.
- **Computação Científica:** Ideal para construir modelos matemáticos e soluções numéricas.

- **Desenvolvimento Geral:** Pode ser aplicada no desenvolvimento de aplicações web, desktop e outras áreas. Em resumo, Julia é uma linguagem versátil que combina alta performance com uma sintaxe amigável, tornando-a uma escolha popular entre cientistas de dados, desenvolvedores e entusiastas da programação.

Visão geral sobre linguagens de programação: Qual a melhor linguagem de programação? Veja https://www.youtube.com/watch?v=DjUB-yVWT2A.

Para iniciantes, recomendamos o curso EAD gratuito no site Goggo dot jl:

https://www.youtube.com/watch?v=0oChN11wf_4

A.1.2 Instalação e execução

Para versões `Linux Ubuntu` ou `Debian`, basta utilizar o comando a seguir e esperar a instalação:

```
> sudo apt install julia -y
```

Para versões `Windows`, utilize o Prompt de comando `PowerShell` para inserir o comando de instalação a seguir:

```
> winget install julia -s msstore
```

Execute o modo interativo da linguagem `Julia` simplesmente digitando o nome da linguagem seguida de Enter:

```
> julia
```

Dessa forma, abre-se o interpretador interativo da linguagem `Julia`. Pode-se então, como primeira interação, realizar uma operação matemática:

```
julia> 2 + 3
5
```

No site oficial da linguagem Julia estão disponíveis para *download* os interpretadores para os principais sistemas operacionais, `Linux`, `Mac OS` e `Windows`.

Além disso, no GitHub do professor Daniel, há um tutorial de instalação para a versão Windows do intepretador de linguagem `Julia` juntamente com o ambiente `Jupyter Notebook` e uma introdução à linguagem de programação com exemplos.

A.1.3 Usando Julia

O uso da linguagem `Julia` pode ser feito de três formas básicas:

- usando um **console Julia** de modo interativo;

- executando um código `codigo.jl` no console `Julia`;
- executando um código `Julia codigo.jl` diretamente no terminal;

Execução no terminal de um código salvo em .jl

Para se executar um código diretamente no terminal de comando do sistema operacional, basta escrevermos o código que desejamos em um arquivo texto de extensão **.jl**.

Dessa forma, por exemplo, salvaremos o seguinte código num arquivo chamado **ola.jl**:

```
1  println("Hello world!!")
```

Após o salvamento, estando o prompt de comando no mesmo diretório do arquivo salvo, basta executarmos o seguinte comando, obtendo-se a resposta que se segue:

```
1 > julia ola.jl
2 Hello world!!
```

Utilização do Console interativo Julia

Para executarmos qualquer comando no console interativo da linguagem `Julia`, iniciaremos o console interativo da seguinte forma:

```
1 > julia
```

Estando o modo interativo rodando no prompt de comando, basta inserirmos o comando desejado e verificar o resultado da saída do comando registrado na linha seguinte do prompt:

```
1 julia> println("Hello world!!")
2 Hello world!!
```

Execução no console de um código salvo em .jl

Para executarmos qualquer um código salvo com extensão **.jl** no console interativo da linguagem `Julia`, iniciaremos o console interativo da seguinte forma:

```
1 > julia
```

Estando o modo interativo rodando no prompt de comando, basta inserirmos o seguinte comando e verificar o resultado da saída do arquivo registrado na linha seguinte do prompt (observe que o arquivo deve estar localizado em ...):

```
julia> include ola.jl
Hello world!!
```

A.2 Elementos da linguagem

`Julia` é uma linguagem de alto nível de tipagem dinâmica, ou seja, uma variável é criada quando um valor é atribuído a ela, não sendo necessário especificar explicitamente cada tipo de variável, `Julia` vai inferir o tipo de cada variável por você. Porém, também é possível especificar a declaração da variável a ser criada, se for desejável.

A.2.1 Variáveis

Variáveis são valores armazenados pelo computador atrelados a um nome específico, para seja possível recuperar ou alterar seu valor posteriormente.

Alguns tipos de variáveis em `Julia`:

- Números inteiros: Int64
- Números reais: Float64
- Matrizes inteiras: Matrix{Int64}
- Matrizes reais: Matrix{Float64}
- Booleanas: Bool
- Strings: String

Por padrão, números são armazenados usando 64 bits, sendo possível aumentar ou reduzir a precisão, utilizando os tipos Int8 ou Int128, por exemplo.

Criamos novas variáveis escrevendo o nome da variável à esquerda e seu valor à direita, e no meio usamos o operador de atribuição =.

Vejamos alguns exemplos:

```
julia> x = 1
1
julia> y = x * 2.0
2.0
```

variaveis com emogi

a variável `x` recebe o valor `int` 1 e, logo após, na segunda linha de comando, a variável `y` recebe o valor `double` 2. Observamos que o símbolo `=` significa o operador de atribuição não o de igualdade. O operador lógico de igualdade no `Julia` é `==`. Veja os seguintes comandos:

```
julia> print(x, "-", y)
1-2.0

julia> typeof(x), typeof(y)
(Int64, Float64)
```

Comentários e continuação de linha de comando são usados como no seguinte exemplo:

```
julia> # Isto é um comentário

julia> x = 1
1

julia> print(x)
1
```

Utilizando Julia como calculadora....

A.3 Repositórios

Tartaruga
OhMyREPL: add OhMyREPL using OhMyREPL
Criar arquivo: .julia/config/startup.jl using OhMyREPL

A.4 Estruturas de ramificação e repetição

A linguagem `Julia` contém estruturas de repetição e ramificação padrões de linguagens estruturadas.

A.4.1 A instrução de ramificação “if”

A instrução “if” permite executar um pedaço do código somente se uma dada condição for satisfeita.

Exemplo A.4.1. Veja o seguinte código Julia:

```
#!/usr/bin/env Julia
# -*- coding:  utf-8 -*-

i = 2
if (i == 1):
    print("Olá!")
elif (i == 2):
    print("Hallo!")
elif (i == 3):
    print("Hello!")
else:
    print("Ça Va!")
```

Qual é a saída apresentada pelo código? Por quê?

Observamos que, em `Julia`, a identação é obrigatória, pois é ela que defini o escopo da instrução.

A.4.2 A instrução de repetição "for"

A instrução `for` permite que um pedaço de código seja executado repetidamente.

Exemplo A.4.2. Veja o seguinte código:

```
for i in range(6):
    print(i)
```

Qual é a saída deste código? Por quê?

Exemplo A.4.3. Veja o seguinte código:

```
import numpy as np
for i in np.arange(1,8,2):
    print(i)
```

Qual é a saída deste código? Por quê?

Exemplo A.4.4. Veja o seguinte código:

```
for i in np.arange(10,0,-3):
  print(i)
```

O que é mostrado no console do Julia?

Exemplo A.4.5. Veja o seguinte código:

```
import numpy as np
for i in np.arange(10,1,-3):
    print(i)
```

O que é mostrado no console do Julia?

A.4.3 A instrução de repetição "while"

A instrução `while` permite que um pedaço de código seja executado repetidamente até que uma dada condição seja satisfeita.

Exemplo A.4.6. Veja o seguinte código Julia:

```
s = 0
i = 1
while (i <= 10):
    s = s + i
    i = i + 1
```

Qual é o valor de `s` ao final da execução? Por quê?

A.5 Funções

Além das muitas funções disponíveis em `Julia` (e os tantos muitos pacotes livres disponíveis), podemos definir nossas próprias funções. Para tanto, existe a instrução `def`. Veja os seguintes exemplos:

Exemplo A.5.1. O seguinte código:

```
def f(x):
    return x + np.sin(x)
```

define a função $f(x) = x + \text{sen } x$.

Observe que $f(\pi) = \pi$. Confirme isso computando:

```
>>> f(np.pi)
```

Exemplo A.5.2. O seguinte código em `Julia`:

```
def h(x,y):
    if (x < y):
        return y - x
    else:
        return x - y
```

define a função:

$$h(x,y) = \begin{cases} y - x & , x < y \\ x - y & , x \geq y \end{cases} \tag{A.1}$$

Exemplo A.5.3. O seguinte código:

```
def J(x):
    y = np.zeros((2,2))
    y[0,0] = 2*x[0]
    y[0,1] = 2*x[1]

    y[1,0] = -x[1]*np.sin(x[0]*x[1])
    y[1,1] = -x[0]*np.sin(x[0]*x[1])

    return y
```

define a matriz jacobiana $J(x_1,x_2) := \frac{(f_1,f_2)}{(x_1,x_2)}$ da função:

$$\boldsymbol{f}(x_1,x_2) = (x_1^2 + x_2^2,\ \cos(x_1 x_2)). \tag{A.2}$$

A.5.1 Operações matemáticas elementares

Em `Julia`, os operadores matemáticos elementares são os seguintes:

```
+ # adição
- # subtração
* # multiplicação
/ # divisão de a por b
\ # divisão de b por a
% # Resto da divisão euclidiana
÷ # Quociente inteiro da divisão euclidiana (alt + 246)
(ou \div + tab)
^ #potenciação
// # Frações
exp() #potenciação de base e
Log() #Logarítimo Neperiano
Log10() #Logarítimo de base 10
```

A.5.2 Funções e constantes elementares

Várias funções e constantes elementares estão disponíveis no pacote módulo Julia math. Por exemplo:

```
julia> π = pi (\pi + tab)
π = 3.1415926535897...

julia> cos(π)
-1.0

julia> exp(1)
2.718281828459045

julia> log(exp(1))
1.0
```

Observamos que `log` é a função logaritmo natural, isto é, $f(x) = \ln(x)$, enquanto que a implementação Julia de $f(x) = \log(x)$ é:

```
julia> log10(10)
1.0
```

Veja mais na documentação Julia.

A.5.3 Operadores lógicos

Em `Julia`, o valor lógico verdadeiro é escrito como `True` e o valor lógico falso como `False`. Temos os seguintes operadores lógicos disponíveis:

```
&& # e lógico
|| # ou lógico
! # negação
== # igualdade
!= # diferente
< # menor que
> # maior que
<= # menor ou igual que
>= # maior ou igual que
```

Veja mais em https://acervolima.com/operadores-em-julia/.

A.6 Matrizes

Em `Julia`, temos um ótimo suporte para computação científica com o pacote numpy. Uma matriz $A = [a_{i,j}]_{i,j=1}^{m,n}$ em Julia é definida usando-se a seguinte sintaxe:

```
julia> A = [
```

```
        a_11  a_12  ...  a_1n ,
        a_21  a_22  ...  a_2n ,
        ⋮
        a_m1  a_m2  ...  a_mn
        ]
```

Exemplo A.6.1. Defina a matriz:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{A.3}$$

Solução. Em `Julia`, digitamos:

```
julia> A = [1 2 3,
            4 5 6]

julia> print(A)
2×3 Matrix{Int64}:
1 2 3
4 5 6
```

$\Diamond$

A seguinte lista contém uma série de funções que geram matrizes particulares:

```
eye # matriz identidade
linspace # vetor de elementos linearmente espaçados
ones # matriz cheia de uns
zeros # matriz nula
```

A.6.1 Obtendo dados de uma matriz

A função `numpy.shape` retorna o tamanho de uma matriz, por exemplo:

```
>>> A = np.ones((3,2))
>>> print(A)
[[ 1. 1.]
 [ 1. 1.]
 [ 1. 1.]]
>>> nl, nc = np.shape(A)
>>> print(nl,nc)
(3, 2)
```

informando que a matriz `A` tem três linhas e duas colunas.

Existem vários métodos para acessar os elementos de uma matriz dada `A`:

- a matriz inteira acessa-se com a sintaxe:

```
A
```

- o elemento da i-ésima linha e j-ésima coluna acessa-se usando a sintaxe:

```
A[i,j]
```

- o bloco formado pelas linhas i_1, i_2 e pelas colunas j_1, j_2 obtém-se usando a sintaxe:

```
A[i1:i2, j1:j2]
```

Exemplo A.6.2. Veja as seguintes linhas de comando:

```
>>> from numpy import random
>>> A = np.random.random((3,4))
>>> A
array([[ 0.39235668, 0.30287204, 0.24379253, 0.98866709],
       [ 0.72049734, 0.99300252, 0.14232844, 0.25604346],
       [ 0.61553036, 0.80615392, 0.22418474, 0.13685148]])
>>> A[2,3]
0.13685147547025989
>>> A[1:3,1:4]
array([[ 0.99300252, 0.14232844, 0.25604346],
       [ 0.80615392, 0.22418474, 0.13685148]])
```

Definida uma matriz A em Julia, as seguintes sintaxes são bastante úteis:

```
A[:,:] toda a matriz
A[i:j,k] os elementos das linhas i até j (exclusive) da k-ésima coluna
A[i,j:k] os elementos da i-ésima linha das colunas j até k (exclusive)
A[i,:] a i-ésima linha da matriz
A[:,j] a j-ésima coluna da matriz
```

Atenção, os índices em `Julia` iniciam-se em 0. Assim, o comando `A[1:3,1:4]` retorna o bloco da matriz A compreendido da segunda à terceira linha e da segunda a quarta coluna desta matriz.

Exemplo A.6.3. Veja as seguintes linhas de comando:

```
>>> B = np.random.random((4,4))
>>> B
array([[ 0.94313432, 0.72650883, 0.55487089, 0.18753526],
       [ 0.02094937, 0.45726099, 0.51925464, 0.8535878 ],
       [ 0.75948469, 0.95362926, 0.77942318, 0.06464183],
       [ 0.91243198, 0.22775889, 0.04061536, 0.14908227]])
>>> aux = np.copy(B[:,2])
>>> B[:,2] = np.copy(B[:,3])
>>> B[:,3] = np.copy(aux)
>>> B
array([[ 0.94313432, 0.72650883, 0.18753526, 0.55487089],
       [ 0.02094937, 0.45726099, 0.8535878 , 0.51925464],
       [ 0.75948469, 0.95362926, 0.06464183, 0.77942318],
       [ 0.91243198, 0.22775889, 0.14908227, 0.04061536]])
```

A.6.2 Operações matriciais e elemento-a-elemento

Em `Julia` com `numpy`, o operador `*` opera elemento a elemento. Por exemplo:

```
>>> A = np.array([[1,2],[2,1]]); print(A)
[[1 2]
 [2 1]]
>>> B = np.array([[2,1],[2,1]]); print(B)
[[2 1]
 [2 1]]
>>> print(A*B)
[[2 2]
 [4 1]]
```

A multiplicação matricial obtemos com:

```
>>> C = A.dot(B)
>>> print(C)
[[6 3]
 [6 3]]
```

Aqui, temos as sintaxes análogas entre operações elemento-a-elemento:

```
+ # adição
- # subtração
* # multiplicação
/ # divisão
^ # potenciação
```

Exemplo A.6.4. Veja as seguintes linhas de comando:

```
>>> A = np.ones((2,2))
>>> A
array([[ 1., 1.],
       [ 1., 1.]])
>>> B = 2 * np.ones((2,2))
>>> B
array([[ 2., 2.],
       [ 2., 2.]])
>>> A*B
array([[ 2., 2.],
       [ 2., 2.]])
>>> A.dot(B)
array([[ 4., 4.],
       [ 4., 4.]])
>>> A/B
array([[ 0.5, 0.5],
       [ 0.5, 0.5]])
```

A.7 Gráficos

Para criar um esboço do gráfico de uma função de uma variável real $y = f(x)$, podemos usar a biblioteca `Julia` mathplotlib.A função `matplotlib.pyplot.plot` faz uma representação gráfica de um conjunto de pontos $\{(x_i, y_i)\}$ fornecidos. Existe uma série de opções para esta função de forma que o usuário pode ajustar várias questões de visualização. Veja a documentação.

Exemplo A.7.1. Veja as seguintes linhas de código:

```
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> def f(x): return x**3 + 1
...
>>> x = np.linspace(-2,2)
>>> plt.plot(x, f(x))
[<matplotlib.lines.Line2D object at 0x7f4f6d153510>]
>>> plt.grid()
>>> plt.show()
```

Resposta dos Exercícios

Recomendamos ao leitor o uso criterioso das respostas aqui apresentadas. Devido a ainda muito constante atualização do livro, as respostas podem conter imprecisões e erros.

E 2.1.1. a) 4; b) 9; c) b^2; d) 7; e) 170; f) 7,125; g) 3,28

E 2.1.2. a) 21,172; b) 5,5; c) 303,25; d) $4,\overline{6}$.

E 2.1.3. $(101,1)_2$.

E 2.1.4. $(11,1C)_{16}$.

E 2.1.5. a) $(12,3\overline{1})_5$; b) $(45,1)_6$.

E 2.1.6. 10,5; $(1010,1)_2$.

E 2.1.7. a) $(100101,001)_2$; b) $(11,4)_{16}$; c) $(11,5)_8$; d) $(9,A)_{16}$.

E 2.1.8. 50; 18.

E 2.2.1.

$$\begin{array}{ll} a)\ 2{,}99792458 \times 10^5 & b)\ 6{,}62607 \times 10^{-34} \\ c)\ 6{,}674 \times 10^{-8} & d)\ 9{,}80665 \times 10^4 \end{array} \tag{2.32}$$

E 2.2.2. No `GNU Octave`, temos:

```
>> printf("%1.7e\n", 299792.458)
2.9979458e+04
>> printf("%1.5e\n", 66.2607)
6.62607e+01
>> printf("%1.3e\n", 0.6674)
6.674e-01
>> printf("%1.5e\n", 9806.65e1)
9.80665e+04
```

E 2.3.1. (a) 1,1; (b) 7,3; (c) $-5,9$.

E 2.3.2. (a) 1,2; (b) 1,2; (c) 2,4; (d) $-2,4$.

E 2.3.3.

Este exercício está sem resposta sugerida. Proponha uma resposta. Veja como em:
https://github.com/livroscolaborativos/CalculoNumerico

E 2.4.1. a) $2^6 + 2^5 + 2^1 = 98$; b) $2^4 + 2^3 + 2^2 + 2^0 = 29$; c) -2^7; d) $-2^7 + 2^6 + 2^5 + 2^1 + 2^0 = -29$; e)$-2^7 + 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 = -1$. Observe que o dígito mais significativo (mais à esquerda) tem peso negativo.

E 2.4.2. a) 25186; b) 7453; c) -7453; d) -1.

E 2.4.3. a) 70; b) -72; c) 72.

E 2.4.4. a) 17990; b) -18248; c) 18248.

E 2.4.5. a) 3,75; b) $-5{,}75$.

E 2.4.7. a) 3,75; b) $-5{,}75$.

E 2.4.8. Devido à precisão finita do sistema de numeração, o laço para quando x for suficientemente grande em comparação a 1 a ponto de x+1 ser aproximado para 1. Isso acontece quando 1 é da ordem do épsilon de máquina em relação a x, isto é, quando $x \approx 2/\%eps$. O tempo de execução fica em torno de 28 anos.

E 2.5.1. a) $\varepsilon_{abs} = 5{,}9 \times 10^{-4}$, $\varepsilon_{rel} = 1{,}9 \times 10^{-2}\%$; b) $\varepsilon_{abs} = \times 10^{-5}$, $\varepsilon_{rel} = \times 10^{-3}\%$; c) $\varepsilon_{abs} = 1$, $\varepsilon_{rel} = 10^{-5}\%$.

E 2.5.2. a) 1,7889; b) 1788,9; c) 0,0017889; d) 0,0045966; e) $2{,}1755 \times 10^{-10}$; f) $2{,}1755 \times 10^{10}$.

E 2.5.3. a) 3270, 3280; b) 42,5, 42,6; c) 0,0000333, 0,0000333.

E 2.5.4. a) 2; b) 2.

E 2.5.5.

$$0{,}1x - 0{,}01 = 12 \tag{2.84}$$
$$0{,}1x = 12 + 0{,}01 = 12{,}01 \tag{2.85}$$
$$x = 120{,}1 \tag{2.86}$$

A resposta exata é 120,1.

E 2.5.6. a) $\delta_{\text{abs}} = 3{,}46 \times 10^{-7}$, $\delta_{\text{rel}} = 1{,}10 \times 10^{-7}$; b) $\delta_{\text{abs}} = 1{,}43 \times 10^{-4}$, $\delta_{\text{rel}} = 1{,}00 \times 10^{-3}$.

E 2.8.1. 2%, deve-se melhorar a medida na variável x, pois, por mais que o erro relativo seja maior para esta variável, a propagação de erros através desta variáveis é muito menos importante do que para a outra variável.

E 2.8.2. 3,2% pela aproximação ou 3,4% pelo segundo método, isto é, $(0{,}96758 \leq I \leq 1{,}0342)$.

E 2.9.1. Quando μ é pequeno, $e^{1/\mu}$ é um número grande. A primeira expressão produz um "overflow" (número maior que o máximo representável) quando μ é pequeno. A segunda expressão, no entanto, reproduz o limite 1 quando $\mu \to 0+$.

E 2.9.2. a) $\frac{1}{2} + \frac{x^2}{4!} + O(x^4)$; b) $x/2 + O(x^2)$; c) $5 \cdot 10^{-4}x + O(x^2)$; d) $\frac{\sqrt{2}}{4}y + O(y^2) = \frac{\sqrt{2}}{4}x + O(x^2)$

E 2.9.3. A expressão da direita se comporta melhor devido à retirada do cancelamento catastrófico em x em torno de 0.

E 2.9.4. Possíveis soluções são:

$$\sqrt{e^{2x}+1}-e^x = \sqrt{e^{2x}+1}-e^x \cdot \frac{\sqrt{e^{2x}+1}+e^x}{\sqrt{e^{2x}+1}+e^x} \tag{2.212}$$

$$= \frac{e^{2x}+1-e^{2x}}{\sqrt{e^{2x}+1}+e^x} = \frac{1}{\sqrt{e^{2x}+1}+e^x} \tag{2.213}$$

e, de forma análoga:

$$\sqrt{e^{2x}+x^2}-e^x = \frac{x^2}{\sqrt{e^{2x}+x^2}+e^x}. \tag{2.214}$$

E 2.9.5. $4{,}12451228 \times 10^{-16}$ J; 0,002%; $0{,}26654956 \times 10^{-14}$ J; 0,002%; $4{,}98497440 \times 10^{-13}$ J; 0,057%; $1{,}74927914 \times 10^{-12}$ J; 0,522%.

E 3.1.1.

Observamos que a equação é equivalente a $\cos(x) - x = 0$. Tomando, então, $f(x) = \cos(x) - x$, temos que $f(x)$ é contínua em $[0, \pi/2]$, $f(0) = 1$ e $f(\pi/2) = -\pi/2 < 0$. Logo, do teorema de Bolzano 3.1.1, concluímos que a equação dada tem pelo menos uma solução no intervalo $(0, \pi/2)$.

E 3.1.2.

No Exercício 3.1.1, mostramos que a função $f(x) = \cos(x) - x$ tem um zero no intervalo $[0, \pi/2]$. Agora, observamos que $f'(x) = -\operatorname{sen}(x) - 1$. Como $0 < \operatorname{sen} x < 1$ para todo $x \in (0, \pi/2)$, temos que $f'(x) < 0$ em $(0, \pi/2)$, isto é, $f(x)$ é monotonicamente decrescente neste intervalo. Logo, da Proposição 3.1.1, temos que existe um único zero da função neste intervalo.

E 3.1.3.

$k \approx 0{,}161228$

E 3.1.5.

Escolhendo o intervalo $[a, b] = [-1{,}841 - 10^{-3}, -1{,}841 + 10^{-3}]$, temos $f(a) \approx 5 \times 10^{-4} > 0$ e $f(b) \approx -1{,}2 \times 10^{-3} < 0$, isto é, $f(a) \cdot f(b) < 0$. Então, o teorema de Bolzano nos garante que o zero exato x^* de $f(x)$ está no intervalo (a, b). Logo, da escolha feita, $|-1{,}841 - x^*| < 10^{-3}$.

E 3.1.6. Basta aplicar as ideias da solução do Exercício 3.1.5.

E 3.2.1. 0,6875

E 3.2.2. Intervalo $(0{,}4, 0{,}5)$, zero 0,45931. Intervalo $(1{,}7, 1{,}8)$, zero 1,7036. Intervalo $(2{,}5, 2{,}6)$, zero 2,5582.

E 3.2.3. a) $x_1 = 1$. b) Dica: como $x_2 = 2$ é raiz dupla, tem-se que $p'(x_2) = 0$.

E 3.2.5. 1,390054; 1,8913954; 2,4895673; 3,1641544; 3,8965468

E 3.2.6. $k\theta = \frac{LP}{2}\cos(\theta)$ com $\theta \in (0, \pi/2)$; 1,030.

E 3.2.7. 19; 23; 26; 0,567143; 1,745528; 3,385630

E 3.2.8. a) 0,623; b) 0,559; c) 0,500; d) 0,300; e) $-0{,}3$; f) -30; g) -30

E 3.2.9. a) 0,0294; b) $2.44e-3$; c) $2.50e-4$; d) $1.09 \cdot 10^{-7}$; e) -10^{-12}; f) -10^{-12}; g) -10^{-12}

E 3.3.1. $-1{,}8414057$

E 3.3.2.
0,7391

E 3.3.3.
Tomemos $x^{(1)} = 1$ como aproximação inicial para a solução deste problema, iterando a primeira sequência a), obtemos:

$$x^{(1)} = 1 \tag{3.79}$$
$$x^{(2)} = \ln\left(\frac{10}{1}\right) = 2{,}3025851 \tag{3.80}$$
$$x^{(3)} = \ln\left(\frac{10}{2{,}3025851}\right) = 1{,}4685526 \tag{3.81}$$
$$\vdots \tag{3.82}$$
$$x^{(21)} = 1{,}7455151 \tag{3.83}$$
$$x^{(31)} = 1{,}745528 \tag{3.84}$$
$$x^{(32)} = 1{,}745528 \tag{3.85}$$

Iterando a segunda sequência b), obtemos:

$$x^{(1)} = 1 \tag{3.86}$$
$$x^{(2)} = 10e^{-1} = 3{,}6787944 \tag{3.87}$$
$$x^{(3)} = 10e^{-3{,}6787944} = 0{,}2525340 \tag{3.88}$$
$$x^{(4)} = 10e^{-0{,}2525340} = 7{,}7682979 \tag{3.89}$$
$$x^{(5)} = 10e^{-7{,}7682979} = 0{,}0042293 \tag{3.90}$$
$$x^{(6)} = 10e^{-0{,}0042293} = 9{,}9577961 \tag{3.91}$$

Este experimento numérico sugere que a iteração a) converge para 1,745528 e a iteração b) não é convergente.

E 3.3.7. $x_1 \approx 1{,}4506619$, $x_2 \approx 4{,}8574864$, $x_3 = 7{,}7430681$.

E 3.3.10.
0.0431266

E 3.4.1. raiz:0,82413, processo iterativo: $x^{(n+1)} = x^{(n)} + \frac{\cos(x) - x^2}{\text{sen}\,(x) + 2x}$

```
>> x=1
>> x=x+(cos(x)-x^2)/(sin(x)+2*x)
>> x=x+(cos(x)-x^2)/(sin(x)+2*x)
>> x=x+(cos(x)-x^2)/(sin(x)+2*x)
>> x=x+(cos(x)-x^2)/(sin(x)+2*x)
```

E 3.4.3. 0,65291864

E 3.4.4. 0,0198679; 0,533890; 0,735412; 1,13237 e 1,38851.

E 3.4.6. -99.99970, -0.3376513; -1.314006.

E 3.4.9.
$x_0 > 1$.

E 3.4.10.

$$x^{(0)} = \text{C.I.} \tag{3.147}$$
$$x^{(n+1)} = x^{(n)}\left(2 - Ax^{(n)}\right) \tag{3.148}$$
$$\tag{3.149}$$

E 3.4.11.

$$x_0 = \text{C.I.} \tag{3.150}$$
$$x^{(n+1)} = x^{(n)}\left(1 - \frac{1}{n}\right) + \frac{A}{nx^{(n)}} \tag{3.151}$$

E 3.4.12.

$$x_0 = \text{C.I.} \tag{3.152}$$
$$x^{(n+1)} = x^{(n)} + \frac{x^{(n)} - Ax^{(n)}}{2} = \frac{(3-A)x^{(n)}}{2} \tag{3.153}$$
$$\tag{3.154}$$

E 3.6.5. Seja $f(x) \in C^2$ um função tal que $f(x^*) = 0$ e $f'(x^*) \neq 0$. Considere o processo iterativo do método das secantes:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})}(x^{(n)} - x^{(n-1)}) \tag{3.202}$$

Esta expressão pode ser escrita como:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} \tag{3.203}$$
$$\tag{3.204}$$
$$= \frac{x^{(n)}\left(f(x^{(n)}) - f(x^{(n-1)})\right) - f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} \tag{3.205}$$
$$= \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} \tag{3.206}$$

Subtraindo x^* de ambos os lados temos:

$$x^{(n+1)} - x^* = \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} - x^* \tag{3.207}$$
$$= \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)}) - x^*\left(f(x^{(n)}) - f(x^{(n-1)})\right)}{f(x^{(n)}) - f(x^{(n-1)})} \tag{3.208}$$
$$= \frac{(x^{(n)} - x^*)f(x^{(n-1)}) - (x^{(n-1)} - x^*)f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} \tag{3.209}$$

Definimos $\epsilon_n = x_n - x^*$, equivalente a $x_n = x^* + \epsilon_n$

$$\epsilon_{n+1} = \frac{\epsilon_n f(x^* + \epsilon_{n-1}) - \epsilon_{n-1}f(x^* + \epsilon_n)}{f(x^* + \epsilon_n) - f(x^* + \epsilon_{n-1})} \tag{3.210}$$

Aproximamos a função $f(x)$ no numerador por

$$f(x^* + \epsilon) \approx f(x^*) + \epsilon f'(x^*) + \epsilon^2 \frac{f''(x^*)}{2} \tag{3.211}$$
$$f(x^* + \epsilon) \approx \epsilon f'(x^*) + \epsilon^2 \frac{f''(x^*)}{2} \tag{3.212}$$

$$\epsilon_{n+1} \approx \frac{\epsilon_n \left[\epsilon_{n-1} f'(x^*) + \epsilon_{n-1}^2 \frac{f''(x^*)}{2}\right] - \epsilon_{n-1}\left[\epsilon_n f'(x^*) + \epsilon_n^2 \frac{f''(x^*)}{2}\right]}{f(x^*+\epsilon_n) - f(x^*+\epsilon_{n-1})} \tag{3.213}$$

$$= \frac{\frac{f''(x^*)}{2}\left(\epsilon_n \epsilon_{n-1}^2 - \epsilon_{n-1}\epsilon_n^2\right)}{f(x^*+\epsilon_n) - f(x^*+\epsilon_{n-1})} \tag{3.214}$$

$$= \frac{1}{2} f''(x^*) \frac{\epsilon_n \epsilon_{n-1}\left(\epsilon_{n-1} - \epsilon_n\right)}{f(x^*+\epsilon_n) - f(x^*+\epsilon_{n-1})} \tag{3.215}$$

Observamos, agora, que

$$\begin{aligned} f(x^*+\epsilon_n) - f(x^*+\epsilon_{n-1}) &\approx \left[f(x^*) + f'(x^*)\epsilon_n\right] - \left[f(x^*) + f'(x^*)\epsilon_{n-1}\right] \\ &= f'(x^*)(\epsilon_n - \epsilon_{n-1}) \end{aligned} \tag{3.216}$$

Portanto:

$$\epsilon_{n+1} \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\epsilon_n \epsilon_{n-1} \tag{3.217}$$

ou, equivalentemente:

$$x^{(n+1)} - x^* \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\left(x^{(n)} - x^*\right)\left(x^{(n-1)} - x^*\right) \tag{3.218}$$

E 3.7.2.
$x > a$ com $a \approx 0{,}4193648$.

E 3.7.3.
$z_1 \approx 0.3252768$, $z_2 \approx 1.5153738$, $z_3 \approx 2.497846$, $z_4 \approx 3.5002901$, $z_j \approx j - 1/2 - (-1)^j \frac{e^{-2j+1}}{\pi}$, $\quad j > 4$

E 3.7.4.
150 W, 133 W, 87 W, 55 W, 6,5 W

E 3.7.5.
a) 42 s e 8 min2 s, b) 14 min56 s.

E 3.7.6.
118940992

E 3.7.7.
7,7 cm

E 3.7.8.
4,32 cm

E 3.7.9.
$(0{,}652919, 0{,}426303)$

E 3.7.10.
7,19% ao mês

E 3.7.11.
4,54% ao mês.

E 3.7.12.
500 K, 700 K em $t = 3\ln(2)$, 26 min, 4 h27 min.

E 3.7.13.
$(\pm 1{,}1101388, -0{,}7675919)$, $(\pm 1{,}5602111, 0{,}342585)$

E 3.7.14.
1,5318075

E 3.7.15.
Aproximadamente 2500 reais por hora.

E 3.7.16.
a) 332,74 K b) 359,33 K

E 3.7.17.
1,2285751, 4,76770758, 7,88704085

E 4.1.1. $p(x) = -3 + 2x + 5x^3$.

E 4.1.2. $p(x) = 0{,}25 + x^2$.

E 4.4.1.
$\int_0^1 P(x)dx = \frac{f(0)+f(1)}{2}$, $\frac{1}{12}\max_{x\in[0,1]}|f''(x)|$