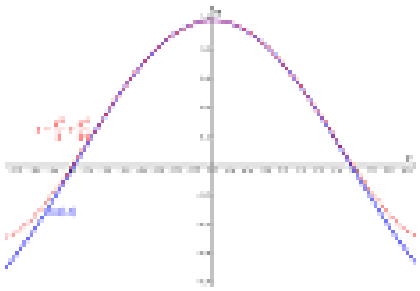
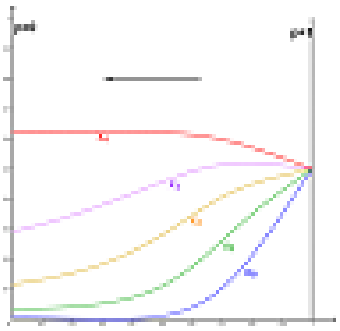
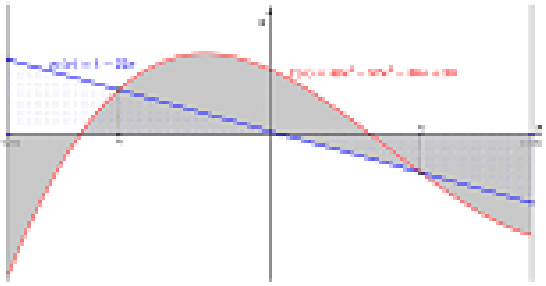
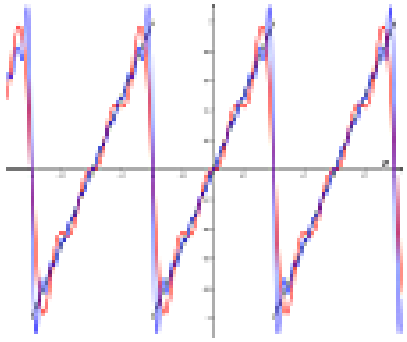
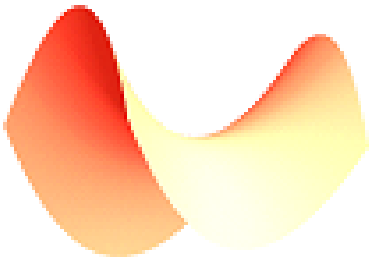
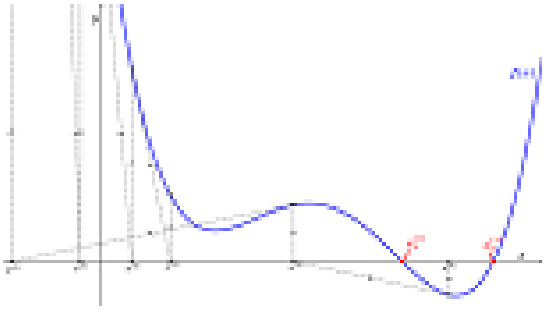
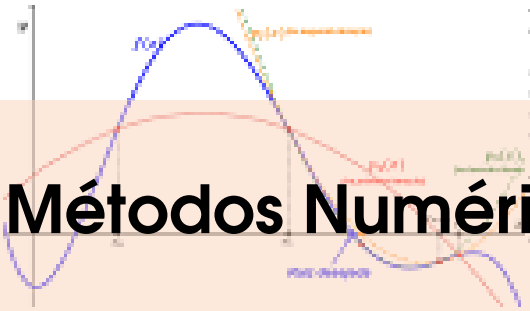
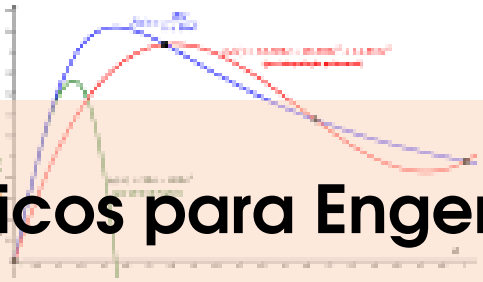
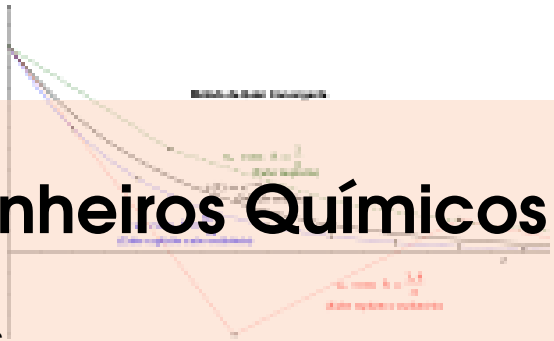
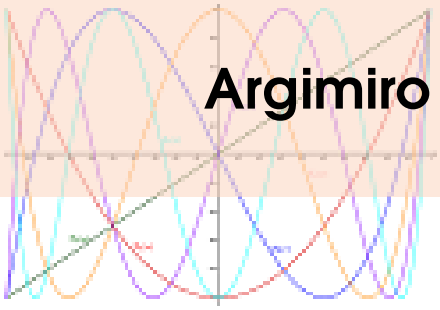
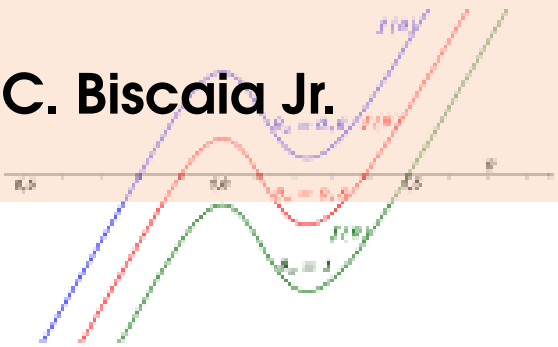
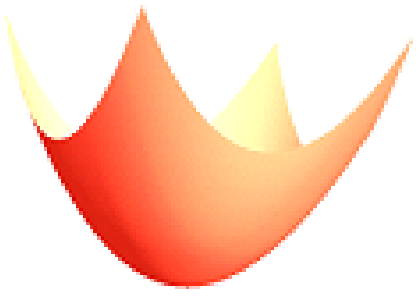
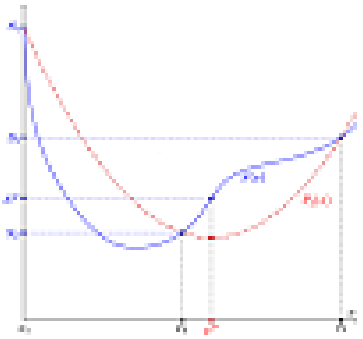
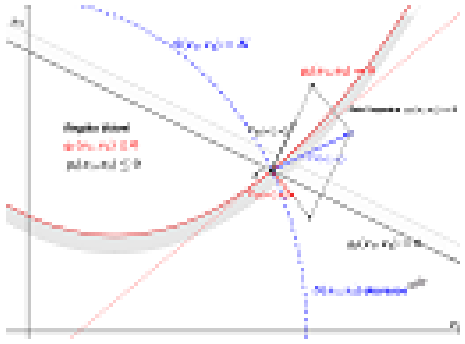
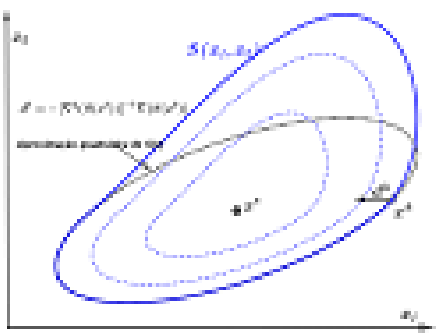
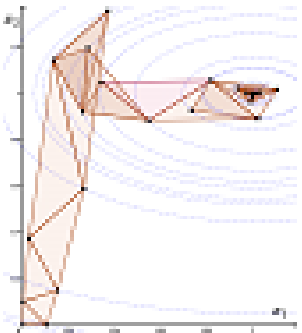
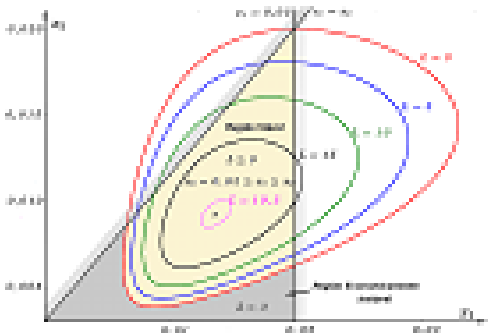










Métodos Numéricos para Engenheiros Químicos

Algoritmos e Aplicações

Argimiro R. Secchi e Evaristo C. Biscaia Jr.



Copyright © 2020 A.R. Secchi e E.C. Biscaia Jr.

PUBLICADO PELOS AUTORES

WWW.PEQ.COPPE.UFRJ.BR

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at `http://creativecommons.org/licenses/by-nc-sa/4.0/`. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

ISBN: 978-65-00-11321-1
Primeira Edição, Março de 2020

Do Argimiro

para suas gatas

Do Evaristo
ao Cícero
e
Bento & Caio

Secchi, Argimiro R.
Métodos numéricos para engenheiros químicos [livro eletrônico] : algoritmos e aplicações / Argimiro R. Secchi, Evaristo C. Biscaia Junior. -- 1. ed. -- Rio de Janeiro : Ed. do Autor, 2020.
PDF

Bibliografia
ISBN 978-65-00-11321-1

1. Bioquímica 2. Engenharia química 3. Indústria alimentícia - Brasil 4. Química 5. Tecnologia I. Jr., Evaristo C. Biscaia. II. Título.

20-47730 CDD-660

Prefácio

Este livro é uma síntese da disciplina de graduação “Métodos Numéricos em Engenharia Química” ministrada há mais de 20 anos no Departamento de Engenharia Química (DEQ) da Escola de Química da UFRJ. Na realidade, a criação da disciplina resultou de negociação, no final da década de 90, com o curso de Matemática Aplicada do Instituto de Matemática da UFRJ no qual se justificou a necessidade desta disciplina ser ministrada no DEQ. Tal justificativa se baseou nos exemplos aplicativos de interesse à engenharia química (EQ), seguindo aproximadamente as diretivas traçadas nos textos pioneiros de Lapidus (1962), "Digital Computation for Chemical Engineers"e de Amundson (1966), "Mathematical Methods in Chemical Engineering: Matrices and Their Application". É importante mencionar que em 1966 a disciplina, então denominada "Cálculo Numérico", foi ministrada pela primeira vez no curso de Engenharia Química na Escola Nacional de Química da Universidade do Brasil pelo Prof. Giulio Massarani, das notas da disciplina resultou o livro "Introdução ao Cálculo Numérico"(Massarani, 1970).

Parte substancial do conteúdo deste livro já se encontrava disponível no site da Internet http://www2.peq.coppe.ufrj.br/Pessoal/Professores/Arge/, elaborada pelos autores, integrando o material da disciplina EQE-358 – Métodos Numéricos em Engenharia Química. O aprofundamento dos conceitos contidos neste material é um dos principais objetivos do presente texto. Durante sua elaboração, surgiram novos e diferentes aspectos de diversos métodos já consagrados, procurando apresentá-los da forma mais pragmática possível, sem excessivo rigor matemático, visando primordialmente aspectos implementacionais.

As implementações dos métodos numéricos apresentados são feitas de forma algorítmica através de pseudo-códigos simples, de fácil compreensão que podem ser realizados por ferramenta computacional de preferência do leitor. Evitou-se ao máximo caracterizar tais procedimentos por meio de softwares comerciais, assegurando com isto a atemporalidade dos mesmos e a perenidade do material apresentado. Muitos exemplos apresentados são de aplicação corrente na EQ e buscam demonstrar ao estudante de graduação de EQ a importância dos métodos numéricos e computacionais neste ramo da engenharia.

Deve-se enfatizar que o aproveitamento do material contido no presente texto é condicionado ao prévio conhecimento do leitor das disciplinas básicas de Cálculo, de Álgebra Linear e de Métodos

Computacionais Básicos. Elementos de álgebra linear são apresentados no Apêndice A para auxiliar na compreensão de métodos numéricos de sistemas de equações.

Por ser um livro escrito a quatro mãos, foi necessário manter uma contínua e constante revisão visando a harmonização e uniformização de seu conteúdo. Mas, felizmente, graças às conexões via Internet foi possível executar as tarefas da maneira mais eficiente possível, apesar da distância entre os locais de trabalho dos autores.

Argimiro Resende Secchi
Evaristo Chalbaud Biscaia Junior

Sumário

1. Introdução

O objetivo deste texto é apresentar e aplicar técnicas e métodos numéricos para a resolução de problemas em processos químicos, bioquímicos e indústrias de alimentos. Os métodos apresentados estão presentes em praticamente todas as ferramentas computacionais usadas pelos engenheiros para sintetizar, analisar, controlar e otimizar tais processos. Espera-se que este texto auxilie no bom uso dessas ferramentas ou no desenvolvimento das mesmas.

Geralmente os métodos numéricos são implementados nessas ferramentas através de uma linguagem de programação, usualmente FORTRAN, C ou C++, e mais recentemente Java e Python. Muitos deles podem ser encontrados em bibliotecas (ou pacotes) numéricos, tais como:

LAPACK (http://www.netlib.org/lapack) – público

BLAS (*Basic Linear Algebra Subprograms*) – público

IMSL (*International Mathematical and Statistical Libraries*) – comercial

NAG (*Numerical Algorithms Group*) – comercial

Mas também estão disponíveis em ambientes dedicados à resolução de problemas genéricos, tais como: MATLAB, OCTAVE, SCILAB, MAPLE, MATHEMATICA, MAXIMA e MATHCAD.

1.1 Sistemas Numéricos

Como a aritmética em calculadoras e computadores emprega apenas números com uma quantidade finita de dígitos, os cálculos são executados com valores aproximados dos números verdadeiros. Para entender essa aritmética, primeiro trataremos da transformação da base decimal para a binária, usada nos processadores numéricos.

Algoritmo 1.1 — Número Inteiro (N) – uso da função $int(x)$.

1. Identificação da maior potência de 2 contida em N, isto é, determinação de n tal que $2^n \leq N < 2^{n+1}$:

$$n = int[log_2(N)]$$

sendo $int(x)$ a parte inteira de x.

2. Determinação dos coeficientes a_i, para $i = 0,1,...,n$, tais que: $N = \sum_{i=0}^{n} a_i 2^i$

$P \leftarrow N$
Para $i = 0,1,...,n,$ faça
$M \leftarrow P/2$
$P \leftarrow int(M)$
$a_i \leftarrow 2(M-P)$

Nota: a operação $2[P/2 - int(P/2)]$ é definida como "P *mod* 2" ou $mod(P,2)$ ou ainda P % 2 e resulta no resto da divisão inteira de P por 2.

Algoritmo 1.2 — Número Inteiro (N) – sem o uso da função $int(x)$.

1. Identificação da maior potência de 2 contida em N, isto é, determinação de n tal que $2^n \leq N < 2^{n+1}$:
$n \leftarrow 0$
$pot \leftarrow 1$
Enquanto $pot \leq N,$ faça
$n \leftarrow n+1$
$pot \leftarrow 2pot$
$pot \leftarrow pot/2$
$n \leftarrow n-1$

2. Determinação dos coeficientes a_i, para $i = 0,1,...,n$, tais que: $N = \sum_{i=0}^{n} a_i 2^i$
$a_n \leftarrow 1$
$M \leftarrow N - pot$
Para $i = 1,2,...,n,$ faça
$pot \leftarrow pot/2$
se $M < pot$ faça $a_{n-i} \leftarrow 0$
senão faça $a_{n-i} \leftarrow 1$ e $M \leftarrow M - pot$

■ **Exemplo 1.1** $N = 97$ (na base decimal)
Algoritmo 1a: $n = int[log_2(97)] = int[6,6] = 6$
Algoritmo 1b:

i	0	1	2	3	4	5	6
M	48,5	24	12	6	3	1,5	0,5
a_i	$a_0 = 1$	$a_1 = 0$	$a_2 = 0$	$a_3 = 0$	$a_4 = 0$	$a_5 = 1$	$a_6 = 1$

Algoritmo 2a:

N	0	1	2	3	4	5	6	7	6
pot	1	2	4	8	16	32	64	128	64

Algoritmo 2b:

i	0	1	2	3	4	5	6
pot	64	32	16	8	4	2	1
M	33	1	1	1	1	1	0
a_{n-i}	$a_6 = 1$	$a_5 = 1$	$a_4 = 0$	$a_3 = 0$	$a_2 = 0$	$a_1 = 0$	$a_0 = 1$

Assim: $97|_{10} = 1100001|_2 = \left(2^6 + 2^5 + 2^0\right)|_{10}$ ■

■ **Exemplo 1.2** Aplicar o Algoritmo 1 para $N = 85$
Algoritmo 1a: $n = int[log_2(85)] = int[6,4] = 6$
Algoritmo 1b:

i	0	1	2	3	4	5	6
M	42,5	21	10,5	5	2,5	1	0,5
a_i	$a_0 = 1$	$a_1 = 0$	$a_2 = 1$	$a_3 = 0$	$a_4 = 1$	$a_5 = 0$	$a_6 = 1$

Assim: $85|_{10} = 1010101|_2 = \left(2^6 + 2^4 + 2^2 + 2^0\right)\big|_{10}$ ■

Generalizando para qualquer base:
$n \leftarrow int[log_{base}(N)]$
$P \leftarrow N$
Para $i = 0, 1, 2, ..., n$, faça
$M \leftarrow P/base$
$P \leftarrow int(M)$
$a_i \leftarrow base(M - P)$

Nota: a operação $base[P/base - int(P/base)]$ é definida como "P *mod* $base$" ou $mod(P, base)$ ou ainda P % $base$ e resulta no resto da divisão inteira de P por $base$.

Algoritmo 1.3 — Número fracionário (α), entre 0 e 1.

1. Especificar o número de dígitos (N_{dig})
2. Determinar os coeficientes b_i, para $i = 1, 2, ..., N_{dig}$, tais que: $\alpha \approx \sum_{i=1}^{N_{dig}} \frac{b_i}{2^i}$

$M \leftarrow 2[\alpha - int(\alpha)]$
Para $i = 1, 2, ..., N_{dig}$, faça
$P \leftarrow int(M)$
$M \leftarrow 2(M - P)$
$b_i \leftarrow P$

■ **Exemplo 1.3** $\alpha = 0,8$
a) $N_{dig} = 8$
b)

i	1	2	3	4	5	6	7	8
M	1,6	1,2	0,4	0,8	1,6	1,2	0,4	0,8
b_i	$b_1 = 1$	$b_2 = 1$	$b_3 = 0$	$b_4 = 0$	$b_5 = 1$	$b_6 = 1$	$b_7 = 0$	$b_8 = 0$

Assim: $0,8|_{10} \approx 0,11001100|_2 = \left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^5} + \frac{1}{2^6}\right)\big|_{10} = 0,796875|_{10}$ ■

■ **Exemplo 1.4** $\alpha = 0,1$
a) $N_{dig} = 8$
b)

i	1	2	3	4	5	6	7	8
M	0,2	0,4	0,8	1,6	1,2	0,4	0,8	1,6
b_i	$b_1 = 0$	$b_2 = 0$	$b_3 = 0$	$b_4 = 1$	$b_5 = 1$	$b_6 = 0$	$b_7 = 0$	$b_8 = 1$

Assim: $0,1|_{10} \approx 0,00011001|_2 = \left(\frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8}\right)\big|_{10} = 0,0976525|_{10}$ ■

A representação da aritmética de ponto flutuante em computação foi normalizada em 1985 pelo IEEE (*Institute for Electrical and Electronic Engineering*) através da Norma 754-1985.

Definições: *bit* – dígito binário
byte – conjunto de 8 bits
word – ou palavra, é a unidade natural do computador para armazenar os números. O tamanho depende da arquitetura do computador, por exemplo:
1 *word* = 32 *bits* = 4 *bytes* ou
1 *word* = 64 *bits* = 8 *bytes*.

Armazenamento de números inteiros:
bit 0: indica o sinal do número, se igual a 0 (zero) número positivo e se igual a 1 número negativo;
bits 1 a 31: codificação do número na base binária quando 1 word = 32 bits.
bits 1 a 63: codificação do número na base binária quando 1 word = 64 bits.

Desta forma o maior número inteiro possível é:

$$2^{31}-1 = 2147483647 \approx 2\times 10^9 \qquad \text{ou} \qquad 2^{63}-1 \approx 9\times 10^{18}$$

■ **Exemplo 1.5** Para ilustrar essa capacidade de armazenamento, faz-se uso do problema proposto por Mordell[1] em 1954, que é a solução da equação Diophantina[2]

$$x^3+y^3+z^3 = n,$$

com x, y e z números inteiros positivos ou negativos e n número inteiro positivo, para o caso específico de $n = 42$. Este problema somente foi resolvido em 2019 por Andrew Booker (Universidade de Bristol) e Andrew Sutherland (MIT), usando 1,3 milhões de horas de computação em rede e mais de 500.000 computadores no sistema *Charity Engine*[3], cuja solução exata é:

$$(-80538738812075974)^3 + 80435758145817515^3 + 12602123297335631^3 = 42.$$

Ao realizar essa operação aritmética em um computador de 32 bits, cuja capacidade máxima de armazenamento de número inteiro é $\vartheta(10^9)$ e os números x, y e z da expressão acima são $\vartheta(10^{16})$, obtém-se um erro absoluto de $1,983\times 10^{35}$. Por outro lado, ao usar um computador de 64 bits, cuja capacidade máxima de armazenamento de números inteiros é $\vartheta(10^{19})$, o valor obtido é exato.

■

Armazenamento de números reais
a) Precisão simples

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

bit 0 (s): indica o sinal do número, se igual a 0 (zero) número positivo e se igual a 1 número negativo;
bits 1 a 8 (c): codificação do expoente (ou característica) de 2 do número (igual ao valor representado em binário menos 127, desta forma o maior expoente é 127 = 11111110 – 127 e o menor expoente é –126 = 00000001 – 127. Esta codificação pode ser lida removendo o bit 1 e somando o seu valor ao bit 8 para os expoentes positivos e a representação complementar para os números negativos, com o primeiro bit representando o sinal negativo. A codificação 11111111 é reservada para infinito e 00000000 para indicar que o número não está normalizado, ou seja, que o número antes da vírgula é zero e não 1);
bits 9 a 31 (f):codificação da mantissa do número (parte fracionária do número no sistema binário).

$r = (-1)^s 2^{c-127}(1+f)$, quando c $\neq 0$
$r = (-1)^s 2^{-127} f$, quando c = 0 (forma não-normalizada)

■ **Exemplo 1.6** Armazenamento (normalizado) de $(3.5)_{10} = (11.1)_2 = (1.11\times 2^1)_2$

0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

■

[1] Louis Joel Mordell (1888-1972).
[2] Diofanto de Alexandria (nascido entre 201 e 214 – falecido entre 284 e 298).
[3] https://phys.org/news/2019-09-sum-cubes-solvedusing-real-life.html

b) Precisão dupla

0	1	2	3	4	5	6	7	8	9	10	11	12	···	63

bit 0 (s): indica o sinal do número, se igual a 0 (zero) número positivo e se igual a 1 número negativo

bits 1 a 11 (c): codificação do expoente de 2 do número (igual ao valor representado em binário menos 1023, desta forma o maior expoente é $1023 = 11111111110 - 1023$ e o menor expoente é $-1022 = 00000000001 - 1023$);

bits 12 a 63 (f): codificação da mantissa do número (parte fracionária do número no sistema binário).

$r = (-1)^s 2^{c-1023}(1+f)$, quando c $\neq 0$

$r = (-1)^s 2^{-1023} f$, quando c = 0 (forma não-normalizada).

Como $63 - 12 + 1 = 52$ dígitos binários correspondem a algo entre 16 a 17 dígitos decimais (até 2^{-52}), então um número representado nesse sistema tem uma precisão de pelo menos 16 dígitos decimais. No caso da precisão simples, esse número cai para 7 dígitos decimais (até 2^{-23}).

■ **Exemplo 1.7** implementar em calculadora ou computador o seguinte algoritmo:

$u \leftarrow 1,0$

Enquanto $u + 1,0 \neq 1,0$, faça

$u \leftarrow 0,5\,u$

$u \leftarrow 2,0\,u$

O valor final em u corresponde à *precisão da máquina*, também conhecido como *epsilon da máquida*. ■

1.2 Erros em Computação

O menor número positivo que pode ser representado na forma normalizada em precisão dupla é:

$$2^{-1022}(1+2^{-52}) \approx 10^{-308}$$

Se os cálculos gerarem números de magnitude menor que este valor, teremos um **erro de *underflow*** (além da capacidade mínima da máquina).

O maior número positivo que pode ser representado na forma normalizada em precisão dupla é:

$$2^{1023}(1+1-2^{-52}) \approx 10^{308}$$

Se os cálculos gerarem números de magnitude maior que este valor, teremos um **erro de *overflow*** (capacidade máxima da máquina excedida).

Erros absolutos e relativos

x: valor exato de um número

x^*: um valor aproximado

Erro absoluto: $EA_x = |x - x^*|$

Erro relativo: $ER_x = \left|\frac{x-x^*}{x}\right| = \left|1 - \frac{x^*}{x}\right| = \frac{EA_x}{|x|}$, para $x \neq 0$

■ **Exemplo 1.8** Se $EA_x = 0,1$

a) com $x = 2112,9$ tem-se: $x - x^* = \pm 0,1 \rightarrow x* = 2112,8$ ou $x^* = 2113$
e $ER_x = 0,1/2112,9 = 4,732832 \times 10^{-5} \approx 0,005\%$

b) com $x = 5,3$ tem-se: $x^* = 5,2$ ou $x^* = 5,4$
e $ER_x = 0,1/5,3 = 0,018868 \approx 2\%$

■

■ **Exemplo 1.9** Efeito da magnitude do número

a) se $x = 0,3000 \times 10^1$ e $x^* = 0,3100 \times 10^1$ tem-se $EA_x = 0,1$ e $ER_x = 0,03333 \approx 3,33\%$
b) se $x = 0,3000 \times 10^{-3}$ e $x^* = 0,3100 \times 10^{-3}$ tem-se $EA_x = 0,1 \times 10^{-4}$ e $ER_x = 0,03333 \approx 3,33\%$
c) se $x = 0,3000 \times 10^4$ e $x^* = 0,3100 \times 10^4$ tem-se $EA_x = 0,1 \times 10^3$ e $ER_x = 0,03333 \approx 3,33\%$

Ou seja, o erro relativo leva em consideração a magnitude dos valores. ■

Diz-se que o número x^* se aproxima do valor x com t **algarismos significativos corretos (ASC)**, ou *dígitos significativos corretos*, se t é o maior valor inteiro não negativo para o qual:

$$ER_x < 5 \times 10^{-t}$$

Neste ponto é importante diferenciar **precisão** de **acurácia** em computação. A precisão indica o quão perto um número representa aquele número que está sendo representado, independente se este estiver correto ou não. Logo, a precisão está diretamente relacionada com a capacidade finita de armazenamento do computador, conforme discutido na Seção 1.1. A acurácia indica o quão perto um número está do valor verdadeiro do número que está sendo representado. Logo, a acurácia está relacionada com os erros da aproximação numérica (erro de arredondamento, erro de convergência do processo iterativo e erro da aproximação inerente ao método numérico). A Figura 1.1 ilustra a diferença entre acurácia e precisão.

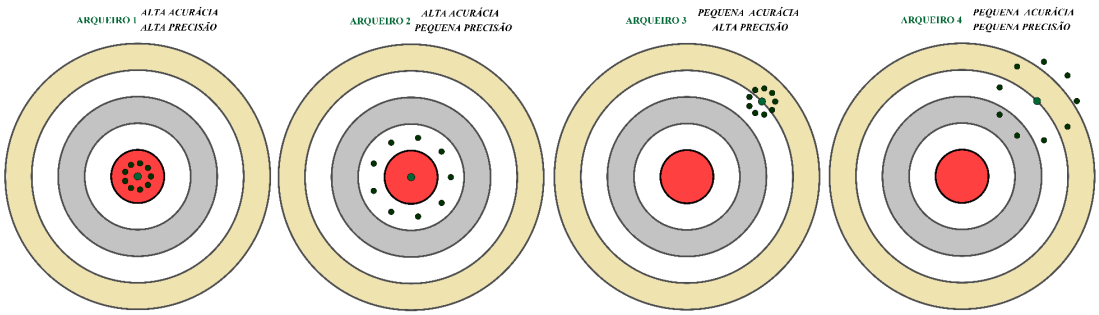


Figura 1.1: Diferença entre acurácia e precisão.

■ **Exemplo 1.10** Exemplo com 4 algarismos significativos
$x = 0,1 \rightarrow ER_x\,|x| = |x - x^*| = EA_x < 5 \times 10^{-5}$
$x = 100 \rightarrow ER_x\,|x| = |x - x^*| = EA_x < 0,05$
$x = 10000 \rightarrow ER_x\,|x| = |x - x^*| = EA_x < 5$ ■

Representando um número x em aritmética de ponto flutuante com t dígitos na base 10:

$$x = f_x \times 10^e + g_x \times 10^{e-t} \text{ com } 0,1 \leq f_x < 1 \text{ e } 0 \leq g_x < 1$$

■ **Exemplo 1.11** $t = 4$ e $x = 234,57$, logo $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1} \rightarrow e = 3, f_x = 0,2345$ e $g_x = 0,7$ ■

Se o número for simplesmente truncado, o valor armazenado de x será: $x^* = f_x \times 10^e$, apresentando:

Erro de Truncamento Absoluto: $EA_x = |g_x| \times 10^{e-t} \leq 10^{e-t}$
Erro de Truncamento Relativo: $ER_x = \left|\frac{g_x \times 10^{e-t}}{f_x \times 10^e + g_x \times 10^{e-t}}\right| < \frac{10^{e-t}}{|f_x \times 10^e|} < \frac{10^{e-t}}{0,1 \times 10^e} = 10^{1-t}$.

Se o número for arredondado, o valor armazenado de x será:

$$x^* = f_x \times 10^e + \begin{cases} 0 & \text{se } g_x < 0,5 \\ 10^{e-t} & \text{se } g_x \geq 0,5 \end{cases},$$

apresentando ***Erros de Arredondamento Absoluto*** **e Relativo** dados por:

i) se $g_x < 0,5$: $EA_x = |g_x| \times 10^{e-t} \leq 0,5 \times 10^{e-t}$
e $ER_x = \left|\frac{g_x \times 10^{e-t}}{f_x \times 10^e + g_x \times 10^{e-t}}\right| < \frac{0,5 \times 10^{e-t}}{|f_x \times 10^e|} < \frac{0,5 \times 10^{e-t}}{0,1 \times 10^e} = 0,5 \times 10^{1-t} = 5 \times 10^{-t}$

ii) se $g_x \geq 0,5$: $EA_x = |g_x \times 10^{e-t} - 10^{e-t}| \leq |0,5 \times 10^{e-t} - 10^{e-t}| \leq 0,5 \times 10^{e-t}$
e $ER_x = \left|\frac{(g_x-1) \times 10^{e-t}}{f_x \times 10^e + g_x \times 10^{e-t}}\right| < \frac{0,5 \times 10^{e-t}}{|f_x \times 10^e|} < \frac{0,5 \times 10^{e-t}}{0,1 \times 10^e} = 0,5 \times 10^{1-t} = 5 \times 10^{-t}$

Resumindo, se: $x = f_x \times 10^e + g_x \times 10^{e-t}$ com $0,1 \leq f_x < 1$ e $0 \leq g_x < 1$, em que t é o número de dígitos, então:

Erros de Truncamento: $\begin{cases} \text{Absoluto} : EA_x < 10^{e-t} \\ \text{Relativo} : ER_x < 10^{1-t} \end{cases}$

Erros de Arredondamento: $\begin{cases} \text{Absoluto} : EA_x < \frac{10^{e-t}}{2} \\ \text{Relativo} : ER_x < \frac{10^{1-t}}{2} \end{cases}$

Erros nas operações algébricas fundamentais

Sejam x e y dois números reais positivos que apresentam erros absolutos máximos a e b, respectivamente. Então os erros numéricos relativos em cada um desses números são:

para x: $p = \frac{a}{x}$
e para y: $q = \frac{b}{y}$.
As seguintes operações aplicadas a esses números são definidas:

i) Soma – o maior valor que a soma $x+y$ pode assumir é: $(x+y)+(a+b)$ e o menor valor é: $(x+y)-(a+b)$, assim:
$(x+y)-(a+b) \leq (x^*+y^*) \leq (x+y)+(a+b)$.
Desse modo: $EA_{x+y} \leq (a+b)$ e $ER_{x+y} \leq \frac{(a+b)}{|x+y|}$

ii) Subtração – o maior valor que a subtração $(x-y)$ pode assumir é: $(x-y)+(a+b)$ e o menor valor é: $(x-y)-(a+b)$, assim:
$(x-y)-(a+b) \leq (x^*-y^*) \leq (x-y)+(a+b)$.
Desse modo: $EA_{x-y} \leq (a+b)$ e $ER_{x-y} \leq \frac{(a+b)}{|x-y|}$.

iii) Produto – o maior valor que o produto (xy) pode assumir é:

$$(x+a)(y+b) = (x+px)(y+qy) = xy(1+p)(1+q) = xy(1+p+q+pq)$$

e o menor valor é: $(x-a)(y-b) = xy(1-p)(1-q) = xy(1-p-q+pq)$
Assim: $xy(1-p-q+pq) \leq x^*y^* \leq xy(1+p+q+pq)$, ou seja:
$xy(pq-p-q) \leq x^*y^* - xy \leq xy(pq+p+q)$. Permitindo identificar que:

$$EA_{xy} \leq |xy|(pq+p+q) \approx |xy|(p+q) \text{ e } ER_{xy} \leq (pq+p+q) \approx p+q$$

iv) Divisão – o maior valor que a divisão (x/y) pode assumir é:

$$\frac{x+a}{y-b} = \frac{x+px}{y-qy} = \frac{x}{y}\left(\frac{1+p}{1-q}\right)\left(\frac{1+q}{1+q}\right) = \frac{x}{y}\left(\frac{1+p+q+pq}{1-q^2}\right)$$

e o menor valor é:

$$\frac{x-a}{y+b} = \frac{x-px}{y+qy} = \frac{x}{y}\left(\frac{1-p}{1+q}\right)\left(\frac{1-q}{1-q}\right) = \frac{x}{y}\left(\frac{1-p-q+pq}{1-q^2}\right).$$

Assim: $\frac{x}{y}\left(\frac{1-p-q+pq}{1-q^2}\right) \leq \frac{x^*}{y^*} \leq \frac{x}{y}\left(\frac{1+p+q+pq}{1-q^2}\right)$, ou seja:

$\frac{x}{y}\left(\frac{1-p-q+pq}{1-q^2}-1\right) \leq \frac{x^*}{y^*}-\frac{x}{y} \leq \frac{x}{y}\left(\frac{1+p+q+pq}{1-q^2}-1\right)$, rearranjando:
$\frac{x}{y}\left(\frac{q^2-p-q+pq}{1-q^2}\right) \leq \frac{x^*}{y^*}-\frac{x}{y} \leq \frac{x}{y}\left(\frac{q^2+p+q+pq}{1-q^2}\right)$, mas:
$q^2-p-q+pq=(p+q)(q-1)$ e $q^2+p+q+pq=(p+q)(q+1)$, logo:
$-\frac{x}{y}\left(\frac{p+q}{1+q}\right) \leq \frac{x^*}{y^*}-\frac{x}{y} \leq \frac{x}{y}\left(\frac{p+q}{1-q}\right)$. Permitindo identificar que:

$$EA_{x/y} \leq \left|\frac{x}{y}\right|\left(\frac{p+q}{1-q}\right) \approx \left|\frac{x}{y}\right|(p+q) \text{ e } ER_{x/y} \leq \frac{p+q}{1-q} \approx p+q.$$

1.3 Problemas Propostos

Problema 1.1 Transforme os números reais abaixo (todos expressos na base decimal) para a base binária, adotando em todos os exemplos 8 dígitos após a vírgula:

a) 168,995889
b) 0,34135
c) 0,021922.

Problema 1.2 Transforme os números binários abaixo para a base decimal:

a) 101101
b) 110101011
c) 0,1100011
d) 0,11111111.

Problema 1.3 Ache um número na base 2 que aproxime o número π com o menor número de dígitos apresentando um erro absoluto não superior a 10^{-3}. Refaça o exemplo para uma aproximação que apresente um erro relativo inferior a 0,01%. Compare e discuta os dois resultados encontrados.

Problema 1.4 Considere que se tem um aparato digital que armazena os números em aritmética de ponto flutuante com quatro dígitos em base decimal. O acumulador (onde são executadas as operações) apresenta precisão dupla (8 dígitos portanto!) e simplesmente trunca os números acumulados. Dados os números: $x = 0,4523 \times 10^4$; $y = 0,2115 \times 10^{-3}$ e $z = 0,2583 \times 10^1$, verifique os resultados das seguintes operações executadas neste aparato e apresente, em cada caso, o erro absoluto e o erro relativo resultante:

a) $x+y+z$
b) $\frac{x}{z}$
c) $x-y$
d) $x-y-z$
e) $\frac{(xy)}{z}$
f) $\left(\frac{x}{z}\right)y$
g) $\left(\frac{y}{z}\right)x$

Compare e discuta os resultados dos itens (e); (f) e (g).

Problema 1.5 Os números $x = 2$ e $y = 1,76$ apresentam, respectivamente, os erros absolutos $a = 0,5$ e $b = 0,1$. Sorteie um valor de x^* e um valor de y^* que estejam contidos entre os valores mínimos e máximos de x e y, respectivamente. Calcule os valores máximos dos erros absolutos e relativos das operações abaixo listadas e verifique se os erros destas mesmas operações aplicadas aos valores sorteados de x e y estão contidos dentro destes limites.

a) $x+y$
b) $x-y$
c) xy
d) $\frac{x}{y}$

e) x^k com $k > 0$
f) x^k com $k < 0$
g) $ln(x)$
h) e^x
i) $\cos(x)$
j) $\text{sen}(x)$
k) $tg(x-1)$.

Problema 1.6 Refaça o Problema 1.5 sabendo-se que no lugar dos erros absolutos são conhecidos os erros relativos de x e y iguais a 20% e 15%, respectivamente.

2. Aproximações de Funções

2.1 Introdução

Há basicamente dois tipos de problemas de aproximações de funções matemáticas:

i) Encontrar uma função *mais simples*, como um polinômio, para aproximar uma função dada de forma explícita;

ii) Encontrar e ajustar a *melhor* função a dados (ou pontos) discretos.

Problemas do tipo *ii*) são abordados no Capítulo 8, com a aplicação do método dos mínimos quadrados.

Existem inúmeras formas para aproximar uma dada função, $f(x)$, por funções *mais simples* ou com propriedades mais convenientes (diferenciação, integração, etc.) tais como:

- aproximação polinomial: $f(x) \approx p_n(x) = \sum_{i=0}^{n} c_i x^i$
- séries de potências: $f(x) \approx f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$
- frações continuadas: $f(x) \approx b_0(x) + \cfrac{a_1(x)}{b_1(x) + \cfrac{a_2(x)}{b_2(x) + \cfrac{a_3(x)}{b_3(x) + \cdots}}}$
- funções racionais: $f(x) \approx \frac{p_n(x)}{q_m(x)} = \frac{\sum_{i=0}^{n} a_i x^i}{\sum_{i=0}^{m} b_i x^i}$
- séries de Fourier[1]: $f(x) \approx a_0 + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \operatorname{sen}(kx)]$

As aproximações polinomiais são tratadas especificamente no Capítulo 3. Portanto, iniciaremos pelas séries de potências.

[1] Jean Baptiste Joseph Fourier (1768-1830).

2.2 Séries de Potências

Se $f(x)$ for uma função contínua com n derivadas contínuas no intervalo $[a,b]$, ou seja, $f(x) \in \mathbb{C}^n[a,b]$ e $\dfrac{d^{n+1}f(x)}{dx^{n+1}}$ existe em $[a,b]$ e $x_0 \in [a,b]$, então

$$f(x) = p_n(x) + R_n(x),$$

em que $p_n(x)$ é o **polinômio de Taylor**[2] de grau n:

$$p_n(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

e

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

é o erro de truncamento (ou resíduo) da série, com $\xi(x) \in (x_0,x)$ se $x > x_0$ ou $\xi(x) \in (x,x_0)$ se $x < x_0$. Essa expressão do erro pode ser entendida através do teorema do valor médio (ou Teorema de Lagrange[3]), ilustrado na Figura 2.1.

Teorema 2.2.1 — Teorema do Valor Médio. Se $f(x)$ é uma função contínua em $[a,b]$ e diferenciável em (a,b), então existe $\xi \in (a,b)$ tal que a reta tangente à curva $f(x)$ no ponto ξ é paralela à reta secante que passa pelos pontos a e b, isto é,

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

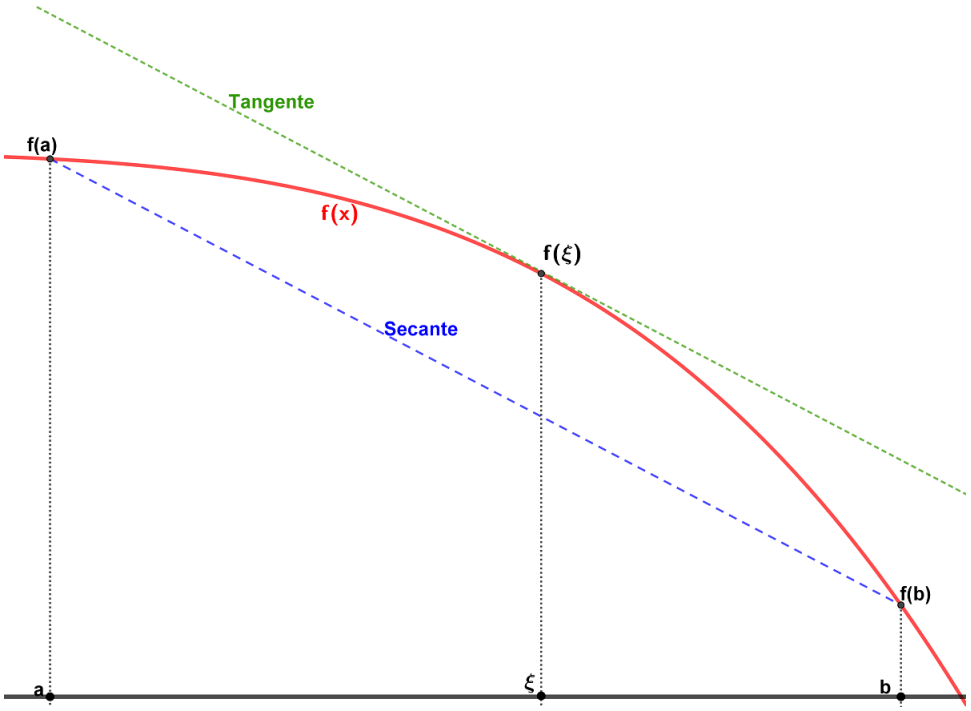


Figura 2.1: Teorema do valor médio.

[2] Brook Taylor (1685-1731).
[3] Joseph-Louis Lagrange (1736-1813).

Aplicando o teorema do valor médio para $a = x_0$ e $b = x$, temos:

$$f'[\xi(x)] = \frac{f(x) - f(x_0)}{x - x_0}.$$

Note que ξ é função de x, pois mantendo o ponto a fixo e variando o ponto b na Figura 2.1, o valor de ξ mudará de posição. Explicitando o valor da função $f(x)$ resulta em:

$$f(x) = f(x_0) + f'[\xi(x)](x - x_0).$$

Comparando essa expressão com a $f(x)$ aproximada pelo Polinômio de Taylor com $n = 0$:

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

obtemos

$$R_0(x) = f(x) - p_0(x) = f'[\xi(x)](x - x_0).$$

Da mesma forma, aplicando o teorema do valor médio para $f'(x)$ obtemos, após integração, a expressão para $R_1(x)$ e assim sucessivamente.

O erro da aproximação pelo Polinômio de Taylor pode ser calculado de três maneiras:

1) $R_n(x) = f(x) - p_n(x)$

2) $R_n(x) = \sum_{k=n+1}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$

3) $R_n(x) = \frac{f^{(n+1)}[\xi(x)]}{(n+1)!}(x - x_0)^{(n+1)}$, $\xi(x) \in (x_0, x)$

Com a primeira forma, a função erro pode ser facilmente construída para todo o domínio. Com a segunda forma seria necessário determinar um número apropriado de termos da série para obter um valor acurado do erro. Finalmente, com a terceira forma seria necessário conhecer *a priori* o valor de $\xi(x)$. Contudo, essa última forma é usada para obter uma estimativa superior do módulo do erro, ao substituir o valor de $\xi(x)$ pelo valor que fornece o maior módulo da $f^{(n+1)}(x)$ no intervalo de interesse, ou seja:

$$|R_n(x)| \leq \left| \frac{f^{(n+1)}[\xi]}{(n+1)!}(x - x_0)^{(n+1)} \right|, \ \xi = \underset{v \in (x_0, x)}{arg\,max} |f^{(n+1)}(v)|.$$

Para uma função $f(x)$ monotônica, o valor de ξ sempre estará em um dos extremos do intervalo de interesse.

Quando $x_0 = 0$, $p_n(x)$ é chamado de **polinômio de Maclaurin**[4].

Quando $n \to \infty$ o polinômio $p_n(x)$ é a **Série de Taylor** (ou **Série de Maclaurin** se $x_0 = 0$).

Um aspecto importante do Polinômio de Taylor é a sua elevada acurácia na vizinhança do ponto x_0, pois o critério para a obtenção dos coeficientes do polinômio $p_n(x)$ é

$$f^{(k)}(x_0) = p_n^{(k)}(x_0), \forall k = 0, 1, \ldots, n.$$

Observe que isto é equivalente a dizer que $R_n^{(k)}(x_0) = 0, \forall k = 0, 1, \ldots, n$, ou seja, x_0 é uma raiz de multiplicidade $n + 1$ da equação $R_n(x) = 0$. Inclusive, com essa informação, é possível também mostrar outro caminho para a obtenção da expressão do erro de truncamento. Fatorando a raiz x_0 temos que $R_n(x) = q(x)(x - x_0)^{(n+1)}$, cuja função $q(x)$ pode ser determinada com a proposição da seguinte função

$$Q(t) = f(t) - p_n(t) - q(x)(t - x_0)^{(n+1)},$$

[4]Colin Maclaurin (1698-1746).

em que x e x_0 são parâmetros dessa função e a equação $Q(t) = 0$ possui pelo menos $n+2$ raízes, sendo $n+1$ em $t = x_0$ e uma raiz em $t = x$. Portanto, $Q^{(n+1)}(t) = 0$ possui pelo menos uma raiz, que chamaremos de $\xi(x)$ e, como $p_n^{(n+1)}(t) = 0$, por ser um polinômio de grau n, resulta em:

$$q(x) = \frac{f^{(n+1)}[\xi(x)]}{(n+1)!},$$

resgatando a expressão do erro de truncamento.

Por outro lado, por ter origem em uma série de potências centrada em x_0, longe desse ponto a acurácia pode ficar comprometida. Por isso, recomenda-se fazer uma normalização da variável $x \in [a,b]$ para o intervalo $[0,1]$ ou para o intervalo $[-1,1]$:

$$y = \frac{x-a}{b-a}, y \in [0,1]$$

ou

$$y = \frac{2x-(b+a)}{b-a}, y \in [-1,1].$$

■ **Exemplo 2.1** Aproximação da função exponencial $f(x) = e^x$ com $x_0 = 0$

$f^{(k)}(x) = e^x \; \forall k$, assim: $f(0) = f'(0) = f''(0) = \cdots = f^{(n)}(0) = 1$, logo:

$$p_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

e

$$R_n(x) = \frac{e^{\xi(x)} x^{n+1}}{(n+1)!}.$$

A expressão de $R_n(x)$ indica que, neste caso, $p_n(x)$ é uma *boa* aproximação de e^x no domínio $|x| < 1$, conforme pode ser observado na Figura 2.2, pois neste domínio $|R_n(x)| \leq \frac{e}{(n+1)!}$. Por exemplo, com $n = 4$ tem-se: $|R_4(x)| \leq \frac{e}{5!} = 0,022652$.

Por outro lado, se o domínio $|x| > 1$, tanto a aproximação quanto a estimativa do erro residual ficam comprometidas. Por exemplo, se $x \in [0,10]$ e desejamos obter o valor aproximado de $f(x)$ para $x = 5$, com $n = 4$, o resultado é o seguinte: $p_4(5) = 65,3750$ e $R_4(5) = 83,0382$, pois $f(5) = 148,4132$. Ao construir uma estimativa superior para o erro da aproximação, o resultado é um valor superestimado: $|R_4(5)| \leq 3,86 \times 10^3$. Portanto, nesse caso é importante aplicar uma normalização do intervalo (mudança de variável), por exemplo, $y = x/10$, resultando em $f(10y) = (e^{10y}) = (e^y)^{10}$ e, então, construir uma aproximação para $g(y) = e^y$:

$$p_4(y) = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \frac{y^4}{4!}.$$

O valor dessa aproximação em $x = 5$, que corresponde a $y = 0,5$, resulta em $p_4(0,5) = 1,64844$ e $[p_4(0,5)]^{10} = 148,1579$, ou seja, um erro da aproximação de $f(5)$ igual a $0,2552$. O valor de $g(0,5) = 1,64872$, logo $R_4(0,5) = g(0,5) - p_4(0,5) = 2,8377 \times 10^{-4}$ e sua estimativa superior $|R_4(0,5)| \leq 4,29 \times 10^{-4}$ resulta em um valor mais próximo do esperado. É possível também construir uma estimativa do erro da função original $f(x)$, ou seja, de $R(10y) = f(10y) - [p_4(y)]^{10} = f(10y) - [g(y) - R_4(y)]^{10}$, usando a fórmula do binômio de Newton[5]:

$$(x+y)^n = \binom{n}{0} x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \cdots + \binom{n}{n} y^n$$

[5] Sir Isaac Newton (1643-1727).

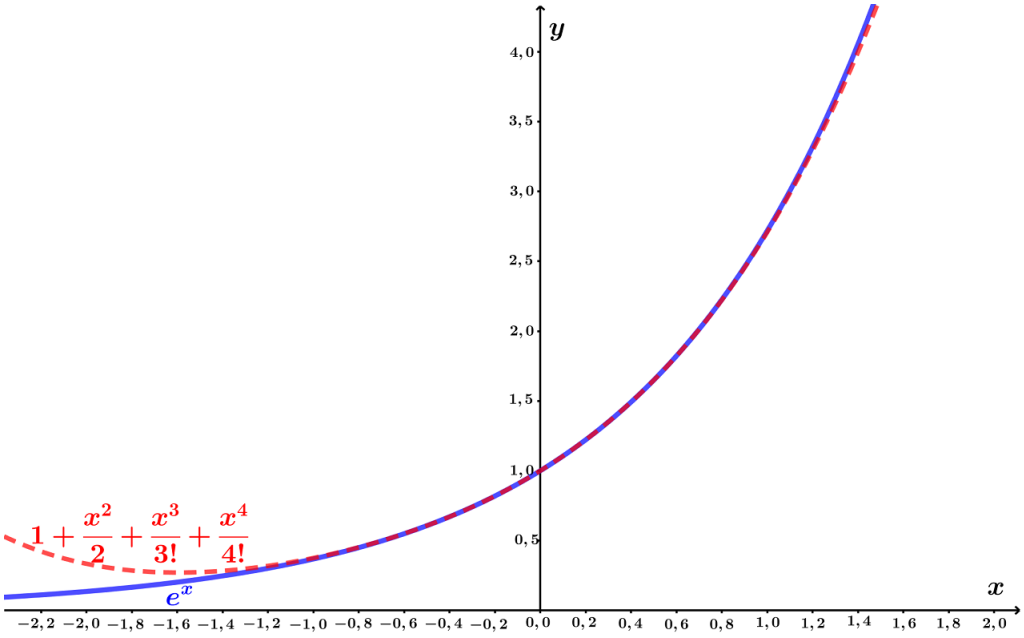


Figura 2.2: Polinômio de Taylor da função exponencial com $n = 4$.

em que $\binom{n}{p} = \dfrac{n!}{p!(n-p)!}$. Dessa forma,

$$R(10y) = f(10y) - [g(y)]^{10} + 10[g(y)]^9 R_4(y) - 45[g(y)]^8 [R_4(y)]^2 + \cdots - [R_4(y)]^{10}.$$

Como o valor de $|R_4(y)| \ll 1$, podemos aproximar essa expressão por: $R(10y) \approx 10[g(y)]^9 R_4(y)$, que avaliada em $y = 0,5$ resulta em $R(5) \approx 0,2554$, ou ainda, usando a estimativa superior para $|R_4(y)|$, resulta em $|R(5)| \leq 0,3865$, que é uma boa estimativa para o erro superior de $f(x)$.

Uma forma de automatizar esse procedimento de normalização do intervalo da função exponencial quando $|x| = \Delta > 1$ é através da busca de m tal que $2^{m-1} < \Delta \leq 2^m$. A variável x é então normalizada segundo:

$$y = \frac{x}{2^m} \Rightarrow e^x = \left[\exp\left(\frac{x}{2^m}\right)\right]^{2^m} = (e^y)^{2^m}.$$

A expansão em série de Taylor de e^y é então efetuada até o termo de grau n tal que

$$|R_n(y)| = \frac{\left|e^{\xi(y)} y^{n+1}\right|}{(n+1)!} \leq \frac{e}{(n+1)!} \leq \delta,$$

sendo δ o critério de convergência. Resultando em: $e^y \approx p_n(y) = \sum_{k=0}^{n} T_k(y)$, em que: $T_k(y) = \dfrac{y^k}{k!}$.

Uma forma *recursiva* de computar o termo T_k é através da *recorrência*:

$$T_k(y) = \frac{y}{k} T_{k-1}(y) \text{ para } k = 1, 2, \ldots, n \text{ com } T_0(y) = 1.$$

Procedimento semelhante é aplicado para computar $p_n(y)$:

$$p_i(y) = \sum_{k=0}^{i} T_k(y) \Rightarrow p_i(y) = p_{i-1}(y) + T_i(y) \text{ para } i = 1, 2, \ldots, n \text{ com } p_0(y) = 1.$$

Outra forma de computar o valor de n é através da verificação da razão: $q_n = \left|\frac{T_n}{p_n}\right|$, sendo esse o valor de n quando se verifica pela *primeira vez* que $q_n \leq \delta$.

A expansão em série de Taylor de e^y, até o valor de n que satisfaz ao critério de convergência considerado, é então efetuada e o valor de e^x é computado através de m potenciações sucessivas de acordo com a fórmula recursiva:
$S_0 = p_n(y)$, $S_k = (S_{k-1})^2$ para $k = 1, 2, \ldots, m$ sendo $e^x \approx S_m$.

Por exemplo, se $x \in [0,10] \longrightarrow \Delta = 10$ logo: $8 = 2^3 < \Delta < 16 = 2^4 \Rightarrow m = 4$ e $y = \frac{x}{16}$, considerando $\delta = 10^{-6}$ tem-se $n = 9$ pois $\frac{e}{9!} = 7{,}49 \times 10^{-6} > \delta$ e $\frac{e}{10!} = 0{,}75 \times 10^{-6} < \delta$, assim: $e^y \approx p_9(y) = \sum_{k=0}^{9} T_k$.

Efetuam-se a seguir quatro potenciações sucessivas: $S_0 = p_9(y)$, $S_1 = S_0^2$, $S_2 = S_1^2$, $S_3 = S_2^2$ e $S_4 = S_3^2$ sendo $e^x \approx S_4$.

A seguir é apresentado um *pseudo-código* de implementação do algoritmo numérico de expansão em série de Taylor de e^x com a especificação de δ.

$y \longleftarrow x$
$m \longleftarrow 0$
$n \longleftarrow 0$
$T \longleftarrow 1$
$S \longleftarrow 1$

Enquanto $|y| \geq 1$, faça
$m \longleftarrow m+1$
$y \longleftarrow \frac{y}{2}$

Faça
$n \longleftarrow n+1$
$T \longleftarrow \frac{y}{n}T$
$S \longleftarrow S+T$
enquanto $\left|\frac{T}{S}\right| > \delta$

Para $j = 1, 2, \ldots, m$, faça
$S \longleftarrow S^2$

No final do algoritmo o valor aproximado de e^x está armazenado na variável S e n é o valor do grau da aproximação $p_n(y)$.

Na Tabela 2.1 são apresentados os valores numéricos resultantes da aplicação do algoritmo com: $x = -2{,}5$ e $\delta = 10^{-4}$.

■

■ **Exemplo 2.2** Aproximação da função $f(x) = \cos(x)$ com $x_0 = 0$ e $n = 4$.

$f'(x) = -\operatorname{sen}(x)$, $f''(x) = -\cos(x)$, $f'''(x) = \operatorname{sen}(x)$, $f^{(4)}(x) = \cos(x)$, $f^{(5)}(x) = -\operatorname{sen}(x)$.
Logo: $f'(0) = 0$, $f''(0) = -1$, $f'''(0) = 0$, $f^{(4)}(0) = 1$ e

$$p_4(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} \quad \text{e} \quad R_5(x) = -\operatorname{sen}(x)\frac{x^5}{120}.$$

A Figura 2.3 ilustra essa aproximação, mostrando a boa qualidade próxima ao centro x_0 e dentro do intervalo $[-1, 1]$.

Tabela 2.1: Valores numéricos resultantes da aplicação do algoritmo de expansão em série de Taylor de e^x com $x = -2,5$ e $\delta = 10^{-4}$

passo	y	m	n	j	T	S	$\left\|\frac{T}{S}\right\|$
0	$-2,5$	0	0		1	1	
1	$-1,25$	1	0		1	1	
2	$-0,625$	2	0		1	1	
3	$-0,625$	2	1		$-0,625$	0,375	1,667
4	$-0,625$	2	2		0,19531	0,57031	0,34247
5	$-0,625$	2	3		$-0,04071$	0,52962	0,07683
6	$-0,625$	2	4		0,00636	0,53598	0,01186
7	$-0,625$	2	5		$-0,00079$	0,53519	0,00148
8	$-0,625$	2	6		0,00008	0,53527	0,00015
9	$-0,625$	2	7		$-0,00001$	0,53526	0,00001
10	$-0,625$	2	7	1	$-0,00001$	0,28650	0,00001
11	$-0,625$	2	7	2	$-0,00001$	**0,08208**	0,00001

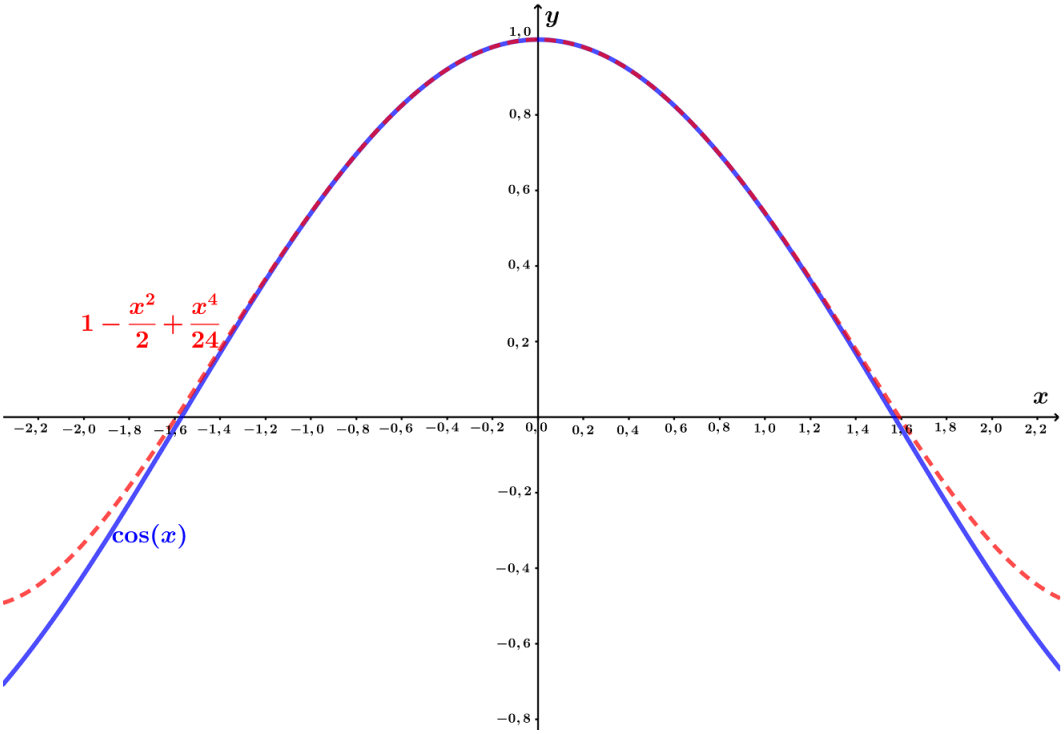


Figura 2.3: Polinômio de Taylor da função cosseno com $n = 4$.

Note que em $x_0 = 0$ todas as derivadas de ordem ímpar de $f(x) = \cos(x)$ são nulas e as derivadas de ordem par satisfazem a $f^{(2n)}(0) = (-1)^n$ permitindo expressar:

$$\cos(x) \approx p_{2n}(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{2k}}{(2k)!}.$$

Sugerido pelo fato dessa expansão conter apenas potências pares de x, propõe-se a mudança de

variável $u = x^2$, resultando em:

$$\cos(x) \approx p_n(u) = \sum_{k=0}^{n} \frac{(-1)^k u^k}{(2k)!} = \sum_{k=0}^{n} T_k(u) \quad \text{sendo} \quad u = x^2 \quad \text{e} \quad T_k(u) = \frac{(-1)^k u^k}{(2k)!}.$$

A forma *recursiva* do cômputo do termo $T_k(u)$ é expressa por:
$T_k(u) = -\dfrac{u}{2k(2k-1)} T_{k-1}(u)$ para $k = 1, 2, \cdots, n$ com $T_0(u) = 1$.
Procedimento semelhante é aplicado para computar $p_n(u)$:

$$p_i(u) = \sum_{k=0}^{i} T_k(u) \Rightarrow p_i(u) = p_{i-1}(u) + T_i(u) \text{ para } i = 1, 2, \cdots, n \text{ e } p_0(u) = 1.$$

É importante enfatizar que a variável x em todas expansões em série de potências de funções trigonométricas deve ser expressa em radianos.

Para assegurar a convergência da série de Taylor da função cosseno deve-se reescalar a variável x (considerada sempre como positiva, sem perda de generalidade) de modo que $x^* \leq 1$, para isto aplicam-se dois procedimentos sequenciais:

1. cálculo do número de *voltas* completas que x contém, isto é: calcular N tal que $2N\pi < x \leq 2(N+1)\pi$, a seguir, subtrai-se de x as *voltas* completas segundo: $x_{novo} = x - 2N\pi$;
2. normalização de x_{novo} que assegure um valor $\leqslant 1$, para isto calcula-se o número inteiro m tal que: $(m-1) < x_{novo} \leq m$, a seguir, reescala-se x_{novo} segundo: $x^* = \dfrac{x_{novo}}{m}$.

$$\cos(x^*) \approx p_{2n}(x^*) = \sum_{k=0}^{n} \frac{(-1)^k (x^*)^{2k}}{(2k)!} = \sum_{k=0}^{n} T_k(x^*)$$

Baseado na expressão:

$$|R_{2n}(x^*)| = \frac{\left|\text{sen}[\xi(x^*)](x^*)^{2n+1}\right|}{(2n+1)!} \leq \frac{(x^*)^{2n+1}}{(2n+1)!} < \frac{(x^*)^{2n}}{(2n)!} = |T_n(x^*)|, \text{ pois } x^* \leq 1.$$

Deste modo, se $|T_n(x^*)| \leq \delta$ tem-se $|R_{2n}(x^*)| < \delta$.

Duas observações são pertinentes aos procedimentos anteriores:

1. o valor de $\cos(x)$ é igual ao valor de $\cos(x_{novo})$ pois x e x_{novo} são arcos congruentes;
2. o valor da expansão de $\cos(x^*)$ é na realidade a expansão de $\cos\left(\dfrac{x}{m}\right)$, ou seja: $\cos(x) = \cos(mx^*)$.

O cômputo do cosseno múltiplo de um arco pode ser efetuado de acordo com o procedimento recursivo descrito a seguir.

$$\cos[(k+1)\theta] = \cos\theta\cos(k\theta) + \text{sen}\,\theta\,\text{sen}(k\theta)$$

$$\cos[(k-1)\theta] = \cos\theta\cos(k\theta) - \text{sen}\,\theta\,\text{sen}(k\theta)$$

Somando as duas equações acima, resulta:

$$\cos[(k+1)\theta] + \cos[(k-1)\theta] = 2\cos\theta\cos(k\theta), \quad \text{ou seja:}$$

$$\cos[(k+1)\theta] = 2\cos\theta\cos(k\theta) - \cos[(k-1)\theta] \text{ para } k = 1, 2, \cdots$$

Definindo $T_k = \cos(k\theta) \Rightarrow T_0 = 1$ e $T_1 = \cos\theta$, tem-se o procedimento recursivo:

$$T_{k+1} = 2T_1T_k - T_{k-1} \text{ para } k = 1, 2, \cdots \text{ com } T_0 = 1 \text{ e } T_1 = \cos\theta.$$

A seguir é apresentado um *pseudo-código* de implementação do algoritmo numérico de expansão em série de Taylor de $\cos x$ com a especificação de δ.

$x \longleftarrow |x|$
$x \longleftarrow x - int\left(\frac{x}{2\pi}\right) \cdot 2\pi$
$m \longleftarrow \text{ceil}(x)$
$u \longleftarrow \left(\frac{x}{m}\right)^2$
$t \longleftarrow 1$
$S \longleftarrow 1$
$n \longleftarrow 0$

Enquanto $|t| \geq \delta$, faça
$n \longleftarrow n+1$
$t \longleftarrow -\frac{u}{2n(2n-1)}t$
$S \longleftarrow S+t$

$T_0 \longleftarrow 1$ e $T_1 \longleftarrow S$

Se $m > 1$ então para $k = 1, 2, \cdots, m-1$ faça
$T_{k+1} \longleftarrow 2T_1T_k - T_{k-1}$

No final do algoritmo o valor aproximado de $\cos(x)$ está armazenado na variável T_m e $2n$ é o valor do grau da aproximação de $\cos(x^*)$.

Na Tabela 2.2 são apresentados os valores numéricos resultantes da aplicação do algoritmo com $x = -65$ e $\delta = 10^{-4}$.

Para iniciar o procedimento deve-se inicialmente reescalar o valor de x, a determinação de x^*, para isto há três pré-etapas:

1. consideração apenas de valores positivos de x, deste modo $|x|$ substitui o valor de x, no exemplo: $x = |-65| = 65$;
2. cômputo do número de voltas completas contidas em $|x|$ para isto calcula-se $N = int\left(\frac{x}{2\pi}\right)$ e, a seguir, subtrai-se de $|x|$ o valor $2N\pi$ resultando em $x_{novo} = |x| - 2N\pi$, no exemplo: $\frac{x}{2\pi} = 10.34507 \Rightarrow N = 10$ e $x_{novo} = 65 - 20\pi = 2,16815$;
3. cômputo do *próximo* inteiro m contido em $x_{novo} : m = \text{ceil}(x_{novo})$. A seguir, o valor de x_{novo} é redimensionado de acordo com $x^* = \frac{x_{novo}}{m}$, no exemplo: $m = \text{ceil}(2,16815) = 3, x^* = \frac{2,16815}{3} = 0,72272$ e $u = (x^*)^2 = 0,52232$.

Tabela 2.2: Valores numéricos resultantes da aplicação do algoritmo de expansão em série de Taylor de $\cos(x)$ com: $x = -65$ e $\delta = 10^{-4}$

passo	1	2	3	4	5	6	7	8
n	0	1	2	3	4	5		
t	1	$-0,26116$	$0,01137$	$-1,98.10^{-4}$	$1,85.10^{-6}$	$-1,07.10^{-8}$		
S	1	0,73884	0,75021	0,75001	0,75001	0,75001		
k							1	2
T_{k+1}							0,12504	**-0,56245**

O procedimento numérico para a expansão em série de potências de $\cos(x)$ pode ser também aplicado para a expansão da função seno, para isto basta substituir o argumento de $\cos(x)$ por $\left(x - \frac{\pi}{2}\right)$, isto é: $\text{sen}(x) = \cos\left(x - \frac{\pi}{2}\right)$. ∎

A seguir são apresentadas algumas expansões em séries de potências de diversas funções contínuas com observações pertinentes às implementações numéricas.

- Logaritmo neperiano (primeira versão)
  $\ln(x) \approx \sum_{k=1}^{n} (-1)^{k-1} \frac{(x-1)^k}{k}$ Série convergente no intervalo: $0 < x < 2$.
  Se $x > 2$ utiliza-se o artifício $x = \frac{1}{\left(\frac{1}{x}\right)}$ e a mudança do argumento x por $y = \frac{1}{x}$, resultando em $0 < y < \frac{1}{2}$ e $\ln(x) = -\ln(y) \approx \left[\sum_{k=1}^{n} (-1)^k \frac{(y-1)^k}{k}\right]$.

- Logaritmo neperiano (segunda versão)
  $\ln(x) \approx 2 \sum_{k=1}^{n} \frac{1}{(2k-1)} \left(\frac{x-1}{x+1}\right)^{2k-1}$
  Como $-1 < \left(\frac{x-1}{x+1}\right) < 1$, $\forall x \in \mathbb{R}^+$, esta expansão é sempre convergente não sendo necessária a mudança do argumento.

- Potenciação genérica de um número real positivo
  $x^q \approx \sum_{k=0}^{n} c_k (x-1)^k$ Série convergente no intervalo: $0 < x \leq 2$.
  Os coeficientes c_i são determinados recursivamente segundo:
  $c_k = \frac{q-(k-1)}{k} c_{k-1}$ para $k = 1, 2, \cdots, n$ com $c_0 = 1$.
  Se $x > 2$ utiliza-se o artifício baseado em: $x^q = \frac{1}{\left(\frac{1}{x}\right)^q}$ e a mudança do argumento x por $y = \frac{1}{x} \Rightarrow 0 < y < \frac{1}{2}$ e $x^q = \frac{1}{y^q} \approx \frac{1}{\sum_{k=0}^{n} c_k (y-1)^k}$.

- Arco seno
  $\text{arcsen}(x) \approx \sum_{k=0}^{n} c_k \frac{x^{2k+1}}{2k+1}$ sendo $|x| < 1$.
  Os coeficientes c_i são determinados recursivamente segundo:
  $c_k = \frac{2k-1}{2k} c_{k-1}$ para $k = 1, 2, n$ com $c_0 = 1$. Esta expansão é sempre convergente não sendo necessária a mudança do argumento.
  A expansão de $\arccos(x)$ é obtida a partir da propriedade: $\arccos(x) = \frac{\pi}{2} - \text{arcsen}(x)$, assim sendo: $\arccos(x) \approx \frac{\pi}{2} - \sum_{k=0}^{n} c_k \frac{x^{2k+1}}{2k+1}$ sendo $|x| < 1$, os coeficientes c_i são os mesmos da expansão do $\text{arcsen}(x)$.

2.3 Frações Continuadas

As aproximações de funções contínuas por frações continuadas (ou contínuas) surgem como uma alternativa às aproximações por série de potências.

A ideia pode ser entendida pelo exemplo numérico abaixo:

$$3,14159 = 3 + \frac{1}{\frac{1}{0,14159}} = 3 + \frac{1}{7,06260} = 3 + \frac{1}{7 + \frac{1}{\frac{1}{0,06260}}} = 3 + \frac{1}{7 + \frac{1}{15,97552}} = \cdots$$

$$= 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{3 + \cdots}}}}}}}}} \approx 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \frac{1}{3}}}}}}}}}$$

Tais frações são determinadas pelo seguinte procedimento recursivo, no qual x é o número a ser aproximado e n o número de frações:

$a \longleftarrow x$

> Para $i = 0, 1, 2, \cdots, n,$ faça
> $b_i \longleftarrow \lfloor a \rfloor$
> Se $i < n$ então $a \longleftarrow \dfrac{1}{a - b_i}$

$x_{aprox} \longleftarrow b_n$

> Para $i = n-1, n-2, \cdots, 1, 0,$ faça
> $x_{aprox} \longleftarrow b_i + \dfrac{1}{x_{aprox}}$

em que $\lfloor x \rfloor$ é o maior inteiro $\leq x$, geralmente implementado nas linguagens de programação pela função $floor(x)$ ou $int(x)$. No exemplo numérico apresentado $x = \pi$ e $n = 10$. O valor aproximado obtido foi 3,1415926536 e $|x - x_{aprox}| = 4,04 \cdot 10^{-13}$, mostrando o alto grau de acurácia do procedimento.

Para aproximações de funções contínuas $f(x)$ o procedimento de frações continuadas pode ser sumarizado pela expressão:

$$f(x) = b_0(x) + \cfrac{a_1(x)}{b_1(x) + \cfrac{a_2(x)}{b_2(x) + \cfrac{a_3(x)}{b_3(x) + \cfrac{a_4(x)}{b_4(x) + \cfrac{a_5(x)}{b_5(x) + \cfrac{a_6(x)}{b_6(x) + \cdots}}}}}}$$

As formas das funções $a_i(x)$ e $b_i(x)$ para $i = 0, 1, 2, \cdots, n$ não são únicas para aproximar uma dada função e podem ser encontradas em Manuais de Matemática para diversas funções $f(x)$.

A forma recursiva de implementação numérica desse procedimento é dada por:

$f_{aprox}(x) \longleftarrow b_n(x)$

Para $i = n-1, n-2, \cdots, 1, 0,$ faça
$f_{aprox}(x) \longleftarrow b_i(x) + \dfrac{a_{i+1}(x)}{f_{aprox}(x)}$

A seguir são apresentados exemplos de expansões em frações continuadas de algumas funções contínuas com os correspondentes resultados numéricos ilustrativos.

■ **Exemplo 2.3** Função exponencial e^x.

(a) Primeira Versão

$a_i = (-1)^{i+1}x$ para $i = 1, 2, \cdots, n$; $b_0 = 1$, $b_i = \begin{cases} i \text{ para } i = 1, 3, 5, 7, \cdots \\ 2 \text{ para } i = 2, 4, 6, 8, \cdots \end{cases}$, assim:

$$e^x \approx 1 + \cfrac{x}{1 - \cfrac{x}{2 + \cfrac{x}{3 + \cfrac{x}{2 - \cfrac{x}{5 + \cfrac{x}{2 + \cfrac{x}{7 - \cfrac{x}{2 + \cfrac{x}{9 - \cfrac{x}{2 + \cdots}}}}}}}}}}$$

Na Tabela 2.3 apresenta-se resultado final do procedimento para $x = 2$, sendo o valor exato de $e^x = 7,389056$, para diferentes valores de n.

Tabela 2.3: Valores numéricos resultantes da aplicação da primeira versão do cômputo de e^x por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6	7	8	9	10	11	12
f_{aprox}	9	7	7,333333	7,4	7,390244	7,388889	7,389041	7,389058	7,389056	7,389056

(b) Segunda Versão

$a_1 = x$, $a_{i+1} = \dfrac{x^2}{4(4i^2-1)}$ para $i = 1, 2, \cdots, n$; $b_0 = 1$, $b_1 = 1 - \dfrac{x}{2}$, $b_i = 1$ para $i = 2, 3, \cdots, n.$, assim:

$$e^x \approx 1 + \cfrac{x}{1 - \cfrac{x}{2} + \cfrac{\frac{x^2}{4\cdot 3}}{1 + \cfrac{\frac{x^2}{4\cdot 15}}{1 + \cfrac{\frac{x^2}{4\cdot 35}}{1 + \cfrac{\frac{x^2}{4\cdot 63}}{1 + \cfrac{\frac{x^2}{4\cdot 99}}{1 + \cdots}}}}}}$$

Na Tabela 2.4 apresenta-se resultado final da segunda versão do procedimento para $x = 2$, para diferentes valores de n.

Tabela 2.4: Valores numéricos resultantes da aplicação da segunda versão do cômputo de e^x por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6	7
f_{aprox}	7,4	7,388889	7,389058	7,389056	7,389056

Note que nesta segunda versão, o procedimento converge ao valor exato da função com menores valores de n.

■

■ **Exemplo 2.4** Função logaritmo neperiano $\ln(x)$.

(a) Primeira Versão

$a_1 = (x-1)$, $a_i = \left[int(\frac{i}{2})\right]^2 (x-1)$ para $i = 2, 3, \cdots, n$; $b_0 = 0, b_i = i$ para $i = 1, 2, \cdots, n$, assim:

$$\ln(x) \approx \cfrac{(x-1)}{1+\cfrac{(x-1)}{2+\cfrac{(x-1)}{3+\cfrac{4(x-1)}{4+\cfrac{4(x-1)}{5+\cfrac{9(x-1)}{6+\cfrac{9(x-1)}{7+\cfrac{16(x-1)}{8+\cdots}}}}}}}}$$

Na Tabela 2.5 apresenta-se resultado final do procedimento para $x = 2$, sendo o valor exato de $\ln(x) = 0,693147$, para diferentes valores de n.

Tabela 2.5: Valores numéricos resultantes da aplicação da primeira versão do cômputo de $\ln(x)$ por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6	7	8	9	10
f_{aprox}	0,7	0,692308	0,693333	0,693122	0,693152	0,693146	0,693147	0,693147

(b) Segunda Versão

$a_1 = 2z$, $a_i = -(i-1)^2 z^2$ para $i = 2, 3, \cdots, n$; $b_0 = 0, b_i = (2i-1)$ para $i = 1, 2, \cdots, n$, em que $z = \left(\frac{x-1}{x+1}\right)$, assim:

$$\ln(x) \approx \cfrac{2z}{1-\cfrac{z^2}{3-\cfrac{4z^2}{5-\cfrac{9z^2}{7-\cfrac{16z^2}{9-\cfrac{25z^2}{11-\cfrac{36z^2}{13-\cfrac{49z^2}{15-\cfrac{64z^2}{17-}}}}}}}}}$$

Na Tabela 2.6 apresenta-se resultado final da segunda versão do procedimento para $x = 2$, para diferentes valores de n.

Tabela 2.6: Valores numéricos resultantes da aplicação da segunda versão do cômputo de $\ln(x)$ por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6
f_{aprox}	0,693122	0,693146	0,693147	0,693147

Observa-se também neste caso que a segunda versão do procedimento converge ao valor exato da função com menores valores de n.

■

■ **Exemplo 2.5** Função tangente trigonométrica $\mathrm{tg}(x)$.
$a_1 = x$, $a_i = -x^2$ para $i = 2, 3, \cdots, n$; $b_0 = 0$, $b_i = (2i-1)$ para $i = 1, 2, \cdots, n$, assim:

$$\mathrm{tg}(x) \approx \cfrac{x}{1 - \cfrac{x^2}{3 - \cfrac{x^2}{5 - \cfrac{x^2}{7 - \cfrac{x^2}{9 - \cfrac{x^2}{11 - \cfrac{x^2}{13 - \cfrac{x^2}{15 - \cfrac{x^2}{17 - \cdots}}}}}}}}}$$

(*)Este procedimento só pode ser aplicado se: $x \neq \left[\frac{\pi}{2} \pm k\pi\right]$

Na Tabela 2.7 apresenta-se resultado final do procedimento para $x = 2$, sendo o valor exato de $\mathrm{tg}(x) = -2,18504$, para diferentes valores de n.

Tabela 2.7: Valores numéricos resultantes da aplicação do cômputo de $\mathrm{tg}(x)$ por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6	7	8
f_{aprox}	-2,444444	-2,20339	-2,185859	-2,185064	-2,18504	-2,18504

■

■ **Exemplo 2.6** Função arco tangente trigonométrica $\mathrm{arctg}(x)$.
$a_1 = x$, $a_i = (i-1)^2 x^2$ para $i = 2, 3, \cdots, n$; $b_0 = 0$, $b_i = (2i-1)$ para $i = 1, 2, \cdots, n$, assim:

$$\mathrm{arctg}(x) \approx \cfrac{x}{1 + \cfrac{x^2}{3 + \cfrac{4x^2}{5 + \cfrac{9x^2}{7 + \cfrac{16x^2}{9 + \cfrac{25x^2}{11 + \cfrac{36x^2}{13 + \cfrac{49x^2}{15 + \cfrac{64x^2}{17 + \cdots}}}}}}}}}$$

Na Tabela 2.8 apresenta-se resultado final do procedimento para $x = 2$, sendo o valor exato de $\text{arctg}(x) = 1,107149$, para diferentes valores de n.

Tabela 2.8: Valores numéricos resultantes da aplicação do cômputo de $\text{arctg}(x)$ por frações continuadas com $x = 2$ para diferentes valores de n

n	13	14	15	16	17	18
f_{aprox}	1,107156	1,107146	1,107150	1,107148	1,107149	1,107149

Note que neste exemplo a convergência do procedimento é mais lenta que nos exemplos anteriores, demandando um número maior de passos. ■

■ **Exemplo 2.7** Função erro $\text{erf}(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x e^{-t^2} dt$.

$a_1 = -\dfrac{e^{-x^2}}{\sqrt{\pi}}$, $a_i = \dfrac{(i-1)}{2}$ para $i = 2, 3, \cdots, n$; $b_0 = 1, b_i = x$ para $i = 1, 2, \cdots, n$, assim:

$$\text{erf}(x) \approx 1 - \cfrac{\frac{e^{-x^2}}{\sqrt{\pi}}}{x + \cfrac{\frac{1}{2}}{1 + \cfrac{\frac{3}{2}}{x + \cfrac{2}{x + \cfrac{\frac{5}{2}}{x + \cfrac{3}{x + \cfrac{\frac{7}{2}}{x + \cfrac{4}{x + \cdots}}}}}}}}$$

Na Tabela 2.9 apresenta-se resultado final do procedimento para $x = 2$, sendo o valor exato de $\text{erf}(x) = 0,995322$, para diferentes valores de n.

Tabela 2.9: Valores numéricos resultantes da aplicação do cômputo de $\text{erf}(x)$ por frações continuadas com $x = 2$ para diferentes valores de n

n	3	4	5	6	7	8
f_{aprox}	0,995303	0,995327	0,995321	0,995323	0,995322	0,995322

■

2.4 Razão de Polinômios

As aproximações de funções por polinômios obtidos do truncamento das expansões em séries de potências podem levar à característica oscilatória indesejada. Tal comportamento pode ser reduzido com a utilização de funções racionais do tipo razões de polinômios, da forma mostrada abaixo:

$$f(x) \approx r(x) = \frac{s_n(x)}{q_m(x)} = \frac{\displaystyle\sum_{i=0}^{n} a_i (x - x_0)^i}{\displaystyle\sum_{j=0}^{m} b_j (x - x_0)^j}$$

Uma técnica bastante empregada para a obtenção dessas razões é a *Técnica de Aproximação de Padé*[6], fundamentada na condição de a expansão em série de potências de $r(x)$ com centro em $x_0 = 0$ ser idêntica à expansão em série de potências de $f(x)$ com mesmo centro até o grau $N = n+m$. Ou seja, $f^{(k)}(x_0) = r^{(k)}(x_0), \forall k = 0,1,\ldots,N$, o que implica que a função erro, $R(x) = f(x) - r(x)$, tem raiz de multiplicidade $N+1$ em $x = x_0$. Assim, expandindo a função $f(x)$ em torno de $x_0 = 0$ resulta:

$$f(x) \approx p_{n+m}(x) = \sum_{k=0}^{N} c_k x^k \text{ sendo } c_k = \frac{f^{(k)}(0)}{k!}.$$

O mesmo tipo de expansão é aplicada à função $r(x)$:

$$r(x) = \frac{\sum_{i=0}^{n} a_i x^i}{\sum_{j=0}^{m} b_j x^j} \approx p_{n+m}(x) = \sum_{k=0}^{N} d_k x^k \text{ sendo } d_k = \frac{r^{(k)}(0)}{k!}.$$

Note que $d_0 = r(0) = \dfrac{a_0}{b_0}$ implicando no fato de que $b_0 \neq 0$, além disto, como $r(x)$ é a razão de dois polinômios o seu valor permanece inalterado multiplicando o numerador e o denominador por uma mesma constante, escolhendo convenientemente esta constante como sendo igual a $\dfrac{1}{b_0}$ resulta em:

$r(x) = \dfrac{\sum_{i=0}^{n} a_i x^i}{1 + \sum_{j=1}^{m} b_j x^j}$ equivalendo a considerar $b_0 = 1$, sem perda de generalidade.

Para determinar os coeficientes dos dois polinômios de $r(x)$ aplica-se a expansão em série de potências à função $R(x)q_m(x) = f(x)q_m(x) - s_n(x) = f(x)\left(1 + \sum_{j=1}^{m} b_j x^j\right) - \sum_{i=0}^{n} a_i x^i$, e, a seguir, iguala-se a zero todos os coeficientes da expansão até o grau $N = n+m$, pois a função $R(x)q_m(x)$ também tem raiz de multiplicidade $N+1$ em $x = x_0$, dando origem a um sistema linear de equações de dimensão $N+1$. Os coeficientes obtidos nessa expansão possuem a seguinte fórmula geral:

$$\sum_{i=0}^{k} c_i b_{k-i} - a_k = 0, \forall k = 0,1,\ldots,N,$$

em que $b_0 = 1$, $b_j = 0$ para $j > m$ e $a_k = 0$ para $k > n$.

■ **Exemplo 2.8** Para exemplificar o procedimento considera-se $f(x) = e^x$ e $n = m = 2$. Assim, expande-se em série de potências em torno de $x_0 = 0$ a função:

$e^x(1 + b_1x + b_2x^2) - (a_0 + a_1x + a_2x^2) \approx (1 - a_0) + (1 - a_1 + b_1)x + \left(\frac{1}{2} - a_2 + b_1 + b_2\right)x^2 +$

$+ \left(\frac{1}{6} + \frac{b_1}{2} + b_2\right)x^3 + \left(\frac{1}{24} + \frac{b_1}{6} + \frac{b_2}{2}\right)x^4$ dando origem ao sistema linear de 5 equações:

$$\begin{aligned}
1 - a_0 &= 0 \\
1 - a_1 + b_1 &= 0 \\
\frac{1}{2} - a_2 + b_1 + b_2 &= 0 \\
\frac{1}{6} + \frac{b_1}{2} + b_2 &= 0 \\
\frac{1}{24} + \frac{b_1}{6} + \frac{b_2}{2} &= 0
\end{aligned}$$

[6] Henri Eugène Padé (1863-1953).

Cuja solução é: $a_0 = 1, a_1 = \frac{1}{2}, a_2 = \frac{1}{12}, b_1 = -\frac{1}{2}, b_2 = \frac{1}{12}$.

Resultando em $r(x) = \dfrac{1+\frac{x}{2}+\frac{x^2}{12}}{1-\frac{x}{2}+\frac{x^2}{12}} = \dfrac{12+6x+x^2}{12-6x+x^2}$. ∎

A seguir são apresentadas diversos exemplos de razões de polinômios para diferentes funções, apresentando-se também os valores máximos dos módulos dos erros nos intervalos especificados.

Tabela 2.10: Aproximações da função exponencial $f(x) = e^x$ por razões de polinômios

$r(x) = \dfrac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $\lvert x\rvert \leq 1$
$\dfrac{2+x}{2-x}$	0,282
$\dfrac{12+6x+x^2}{12-6x+x^2}$	$4 \cdot 10^{-3}$
$\dfrac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3}$	$2,8 \cdot 10^{-5}$
$\dfrac{1680+840x+180x^2+20x^3+x^4}{1680-840x+180x^2-20x^3+x^4}$	$1,1 \cdot 10^{-7}$

Tabela 2.11: Aproximações da função cosseno trigonométrico $f(x) = \cos(x)$ (após aplicação da mudança de variável $u = x^2$) por razões de polinômios

$r(x) = \dfrac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $\lvert x\rvert \leq 1$
$\dfrac{12-5x^2}{12+x^2}$	$1,8 \cdot 10^{-3}$
$\dfrac{15120-6900x^2+313x^4}{15120+660x^2+13x^4}$	$3,6 \cdot 10^{-7}$
$\dfrac{39251520-18471600x^2+1075032x^4-14615x^6}{39251520+1154160x^2+16632x^4+127x^6}$	$1,27 \cdot 10^{-11}$

Tabela 2.12: Aproximações da função seno trigonométrico $f(x) = \text{sen}(x)$ (após aplicação da aproximação em $\dfrac{f(x)}{x}$ seguida da mudança de variável $u = x^2$) por razões de polinômios

$r(x) = \dfrac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $\lvert x\rvert \leq 1$
$\dfrac{60x-7x^3}{60+3x^2}$	$2,01 \cdot 10^{-4}$
$\dfrac{166320x-22260x^3+551x^5}{166320+5460x^2+75x^4}$	$2,27 \cdot 10^{-8}$
$\dfrac{11511339840x-1640635920x^3+52785432x^5-479249x^7}{11511339840+277920720x^2+3177720x^4+18361x^6}$	$5,67 \cdot 10^{-13}$

Tabela 2.13: Aproximações da função logaritmo neperiano $f(x) = \ln(1+x)$ por razões de polinômios

$r(x) = \frac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $0 \leq x \leq 1$
$\frac{2x}{2+x}$	0,026
$\frac{6x+3x^2}{6+6x+x^2}$	$8,4 \cdot 10^{-4}$
$\frac{60x+60x^2+11x^3}{60+90x+36x^2+3x^3}$	$2,55 \cdot 10^{-5}$
$\frac{420x+630x^2+260x^3+25x^4}{420+840x+540x^2+120x^3+6x^4}$	$7,6 \cdot 10^{-7}$

Tabela 2.14: Aproximações da função tangente trigonométrica $f(x) = \text{tg}(x)$ por razões de polinômios

$r(x) = \frac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $\lvert x\rvert \leq 1$
$\frac{3x}{3-x^2}$	0,057
$\frac{15x-x^3}{15-6x^2}$	$8,4 \cdot 10^{-4}$
$\frac{60x+60x^2+11x^3}{60+90x+36x^2+3x^3}$	$1,85 \cdot 10^{-3}$
$\frac{945x-105x^3+x^5}{945-420x^2+15x^4}$	$3,2 \cdot 10^{-7}$

Tabela 2.15: Aproximações da função arco tangente trigonométrica $f(x) = \text{arctg}(x)$ por razões de polinômios

$r(x) = \frac{s_n(x)}{q_m(x)}$	$\lvert Erro(x)\rvert_{max}$ para $\lvert x\rvert \leq 1$
$\frac{15x+4x^3}{15+9x^2}$	$6,3 \cdot 10^{-3}$
$\frac{945x+735x^3+64x^5}{945+1050x^2+225x^4}$	$1,9 \cdot 10^{-4}$
$\frac{15015x+19250x^2+5943x^3+256x^7}{15015+24255x^2+11025x^4+1225x^6}$	$5,56 \cdot 10^{-6}$

2.5 Séries de Fourier

Nas expansões de funções contínuas $f(x)$, com n derivadas contínuas no intervalo $[a,b]$, por séries de potências adotam-se como *bases* das expansões as sucessivas potências de $(x-x_0)$ e os coeficientes c_i podem ser interpretados como os *componentes* da expansão em relação à base considerada. Uma alternativa bastante conveniente como bases das expansões de funções é a adoção de funções $y_i(x)$ que apresentam a propriedade de *ortogonalidade* no intervalo $[a,b]$ em relação a uma determinada função *peso*: $w(x)$, contínua e não negativa no intervalo. A propriedade de ortogonalidade dessas funções é caracterizada pela expressão:

$$\int_a^b w(x)y_i(x)y_j(x)dx = \begin{cases} 0 \text{ para } i \neq j \\ K_i > 0 \text{ para } i = j \end{cases} = K_i\delta_{ij}, \text{ sendo } \delta_{ij} = \begin{cases} 0 \text{ para } i \neq j \\ 1 \text{ para } i = j \end{cases}$$

O termo δ_{ij} é chamado de delta de Kronecker[7].

Deste modo, propõe-se a aproximação da função $f(x)$, por $f(x) \approx \sum_{i=0}^{n} c_i y_i(x)$. O erro desta aproximação é expresso por: $R_n(x,\mathbf{c}) = f(x) - \sum_{i=0}^{n} c_i y_i(x)$.

Uma forma de avaliar este erro em todo o intervalo $[a,b]$ é através da integral do seu quadrado

$$J(\mathbf{c}) = \int_a^b w(x)R_n^2(x,\mathbf{c})dx \text{ sendo } \mathbf{c} = \begin{bmatrix} c_0 & c_1 & \cdots & c_n \end{bmatrix}^T.$$

Os *melhores* valores dos coeficientes são aqueles que minimizam $J(\mathbf{c})$, devendo satisfazer a: $\frac{\partial J(\mathbf{c})}{\partial c_i} = 0$ para $i = 0, 1, 2, \cdots, n$. Ou seja: $\frac{\partial J(\mathbf{c})}{\partial c_i} = 2\int_a^b w(x)R_n(x,\mathbf{c})\frac{\partial R_n(x,\mathbf{c})}{\partial c_i}dx = 0$.

Como: $\frac{\partial R_n(x,\mathbf{c})}{\partial c_i} = -y_i(x) \Rightarrow \frac{\partial J(\mathbf{c})}{\partial c_i} = -2\int_a^b w(x)[f(x) - \sum_{j=0}^{n} c_j y_j(x)]y_i(x)dx$.

Então: $\frac{\partial J(\mathbf{c})}{\partial c_i} = 0 \Rightarrow \sum_{j=0}^{n}\left[\int_a^b w(x)y_i(x)y_j(x)dx\right]c_j = \int_a^b w(x)y_i(x)f(x)dx$, em vista de:

$\int_a^b w(x)y_i(x)y_j(x)dx = K_i\delta_{ij}$, tem-se $\sum_{j=0}^{n}\left[\int_a^b w(x)y_i(x)y_j(x)dx\right]c_j = K_i c_i$.

Resultando em:

$$c_i = \frac{\int_a^b w(x)y_i(x)f(x)dx}{K_i} = \frac{\int_a^b w(x)y_i(x)f(x)dx}{\int_a^b w(x)y_i^2(x)dx}.$$

Considerando $w(x) = 1$, $a = -1$ e $b = 1$ as funções $\cos(i\pi x)$ para $i = 0, 1, 2, \cdots n, \text{sen}(i\pi x)$ para $i = 1, 2, \cdots n$ ou combinações lineares das mesmas são ortogonais entre si, pois:

$$\int_{-1}^{+1}\cos(i\pi x)\cos(j\pi x)dx = \begin{cases} 0 \text{ para } i \neq j \\ 1 \text{ para } i = j \neq 0 \\ 2 \text{ para } i = j = 0 \end{cases},$$

$$\int_{-1}^{+1}\text{sen}(i\pi x)\,\text{sen}(j\pi x)dx = \begin{cases} 0 \text{ para } i \neq j \\ 1 \text{ para } i = j \neq 0 \\ 0 \text{ para } i = j = 0 \end{cases} \text{ e}$$

$\int_{-1}^{+1}\cos(i\pi x)\,\text{sen}(j\pi x)dx = 0$ em todos os casos.

Quando estas funções são consideradas a expansão:

$$f(x) = a_0 + \sum_{i=1}^{\infty}[a_i\cos(i\pi x) + b_i\,\text{sen}(i\pi x)] \text{ é chamada de } \textbf{Série de Fourier}.$$

Seus coeficientes são determinados pela minimização de:

$$J(\mathbf{a},\mathbf{b}) = \int_{-1}^{+1}\left\{f(x) - a_0 - \sum_{i=1}^{\infty}[a_i\cos(i\pi x) + b_i\,\text{sen}(i\pi x)]\right\}^2 dx.$$

[7] Leopold Kronecker (1823-1891).

Assim: $\dfrac{\partial J(\mathbf{a},\mathbf{b})}{\partial a_i} = 0 \;\; \dfrac{\partial J(\mathbf{a},\mathbf{b})}{\partial b_i} = 0,$ resultando em: $a_0 = \dfrac{\int_{-1}^{+1} f(x)dx}{2} = f_{medio}(x),$
$a_i = \int_{-1}^{+1} f(x)\cos(i\pi x)dx$ e $b_i = \int_{-1}^{+1} f(x)\,\mathrm{sen}(i\pi x)dx$ para $i = 1, 2, \cdots$

Uma propriedade importante das Séries de Fourier é sua periodicidade, pois:
$\cos(i\pi x) = \cos[i\pi(x+2k)]$ e $\mathrm{sen}(i\pi x) = \mathrm{sen}[i\pi(x+2k)] \Rightarrow f(x) = f(x+2k)$ para $k = 0, \pm1, \pm2, \cdots$
Outro aspecto que deve ser enfatizado é a restrição destas expansões ao intervalo $-1 \leq x \leq +1$, valores de x fora deste intervalo repetem o valor da função dentro do intervalo $-1 \leq x \leq +1$, devido à periodicidade da mesma. Tal comportamento pode ser ilustrado pelos exemplos numéricos:
$f(5,45) = f(6-0,45) = f(-0,45)$ e $f(-7,67) = f[-(8-0,33)] = f(0,33)$.

Se a função $f(x)$ for uma função *par* tem-se: $f(x) = f(-x)$, uma vez que a função $\cos(i\pi x)$ é uma função par, o produto $f(x)\cos(i\pi x)$ é também par, no entanto, como a função $\mathrm{sen}(i\pi x)$ é uma função ímpar, o produto $f(x)\,\mathrm{sen}(i\pi x)$ é também ímpar.
Deste modo, se a função $f(x)$ for uma função par tem-se: $a_0 = \int_0^{+1} f(x)dx = f_{medio}(x)$,
$a_i = 2\int_0^{+1} f(x)\cos(i\pi x)dx$, $b_i = 0$ para $i = 1, 2, \cdots$ e a Série de Fourier correspondente só terá os termos de $\cos(i\pi x)$, sendo assim chamada de **Série Cosseno de Fourier** e expressa por:

$$f(x) = a_0 + \sum_{i=1}^{\infty} a_i \cos(i\pi x).$$

Se a função $f(x)$ for uma função *ímpar* tem-se: $f(x) = -f(-x)$, uma vez que a função $\cos(i\pi x)$ é uma função par, o produto $f(x)\cos(i\pi x)$ é ímpar, no entanto, como a função $\mathrm{sen}(i\pi x)$ é uma função ímpar, o produto $f(x)\,\mathrm{sen}(i\pi x)$ é par.
Deste modo, se a função $f(x)$ for uma função ímpar tem-se: $a_0 = 0$, $a_i = 0$, $b_i = 2\int_0^{+1} f(x)\,\mathrm{sen}(i\pi x)dx$ para $i = 1, 2, \cdots$ e a Série de Fourier correspondente só terá os termos de $\mathrm{sen}(i\pi x)$, sendo assim chamada de **Série Seno de Fourier** e expressa por: $f(x) = \sum_{i=1}^{\infty} b_i \,\mathrm{sen}(i\pi x)$

■ **Exemplo 2.9** $f(x) = \begin{cases} +1 \text{ para } 0 < x < 1 \\ -1 \text{ para } -1 < x < 0 \end{cases}$ (função ímpar) então:

$$b_i = 2\int_0^{+1} \mathrm{sen}(i\pi x)dx = 2\frac{1-(-1)^i}{i\pi} = \begin{cases} 0 \text{ para } i = 2, 4, 6, \cdots, 2k \\ \dfrac{4}{\pi i} \text{ para } i = 1, 3, 5, \cdots, 2k-1 \end{cases}$$

$$f(x) = \frac{4}{\pi}\sum_{i=1}^{\infty} \frac{\mathrm{sen}[(2i-1)\pi x]}{2i-1}.$$

■

■ **Exemplo 2.10** $f(x) = |x|$, (função par) então:

$$a_0 = \int_0^{+1} x dx = \frac{1}{2},\; a_i = 2\int_0^{+1} x\cos(i\pi x)dx = 2\frac{(-1)^i - 1}{(\pi)^2 i^2} = \begin{cases} 0 \text{ para } i = 2, 4, 6, \cdots, 2k \\ -\dfrac{4}{(\pi)^2 i^2} \text{ para } i = 1, 3, \cdots, 2k-1 \end{cases}.$$

$$f(x) = |x| = \frac{1}{2} - \frac{4}{(\pi)^2}\sum_{i=1}^{\infty} \frac{\cos[(2i-1)\pi x]}{(2i-1)^2}.$$

■

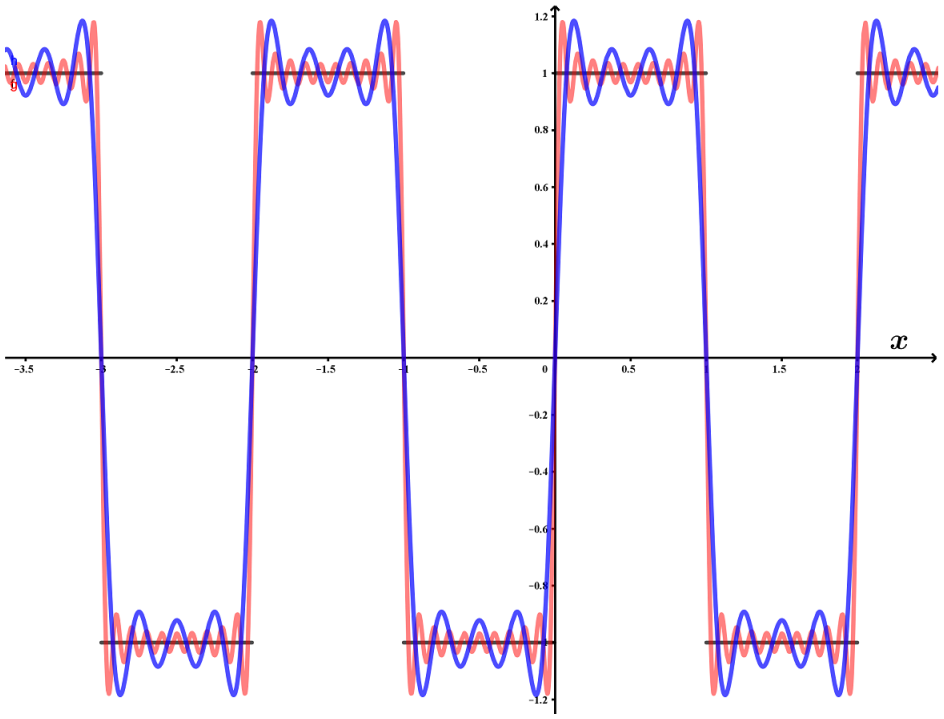


Figura 2.4: Série de Fourier do trem de pulsos do Exemplo 2.9 com $n = 4$ (curva azul) e $n = 10$ (curva vermelha).

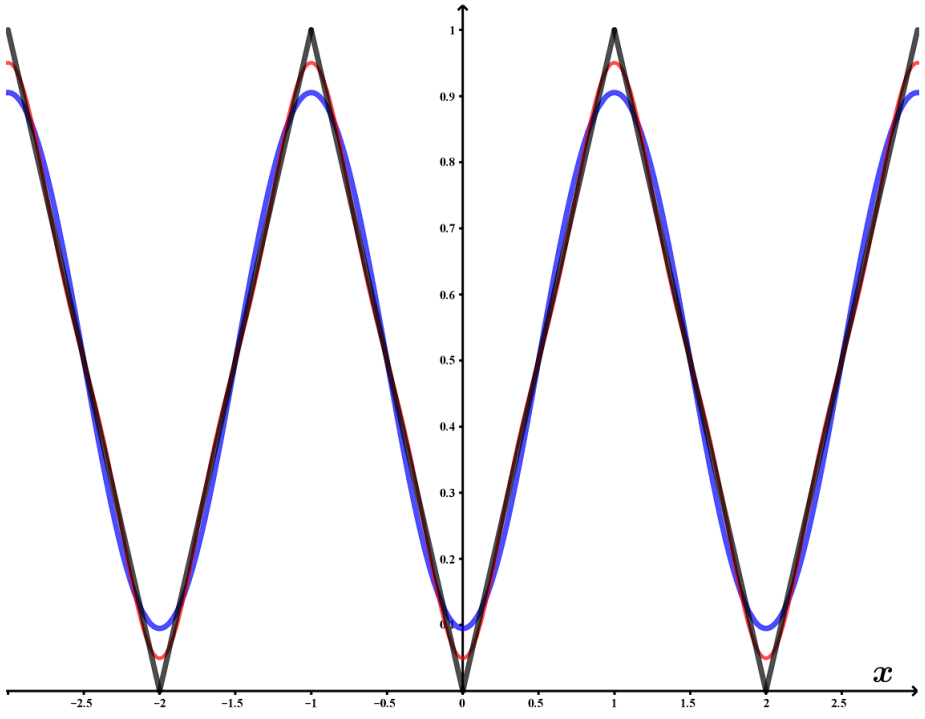


Figura 2.5: Série de Fourier de $|x|$ com $n = 1$ (curva azul) e $n = 4$ (curva vermelha).

■ **Exemplo 2.11** $f(x) = x$, (função ímpar) então: $b_i = 2\int_0^{+1} x\,\mathrm{sen}(i\pi x)dx = 2\dfrac{(-1)^{(i+1)}}{i\pi}$.

$$f(x) = x = \frac{2}{\pi}\sum_{i=1}^{\infty}\frac{(-1)^{i+1}\,\mathrm{sen}(i\pi x)}{i}.$$

■

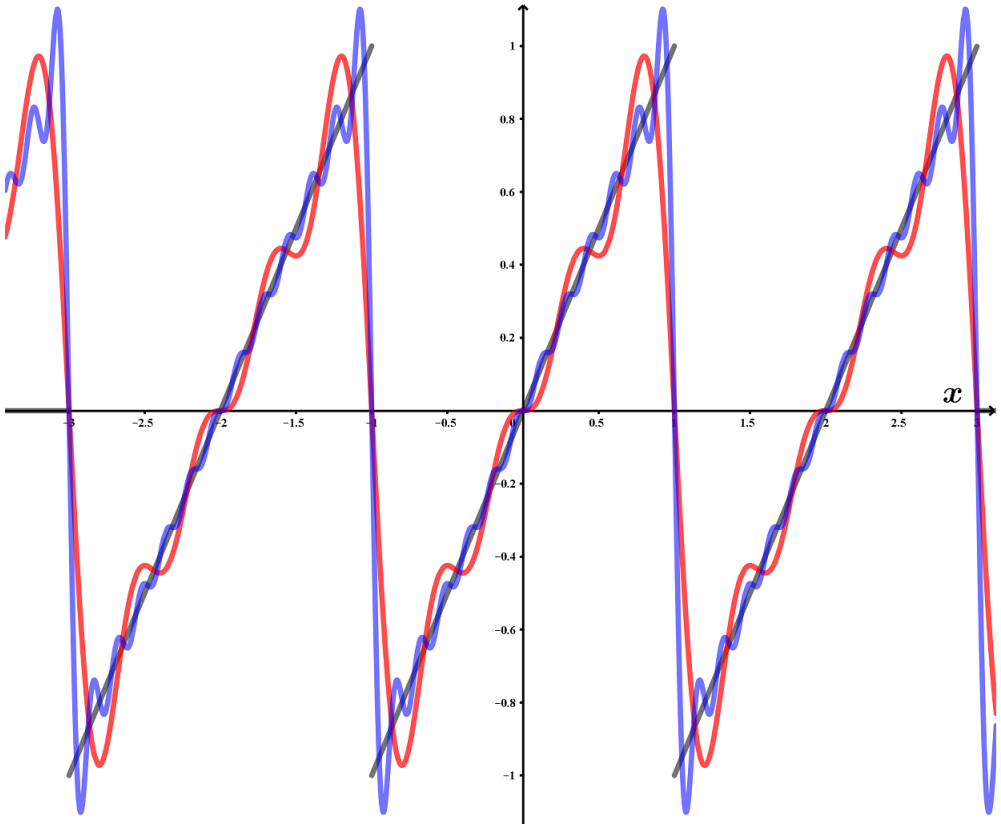


Figura 2.6: Série de Fourier de x com $n = 2$ (curva vermelha) e $n = 12$ (curva azul).

2.6 Problemas Propostos

Problema 2.1 Visando a aproximação da função exponencial para valores negativos elevados de seu argumento que leve em conta que $\lim_{x\to\infty} e^{-x} = 0$, propõe-se a aproximação de e^x pela razão de polinômios com o polinômio do denominador um grau maior do que o grau do polinômio do numerador, isto é $e^x \approx \dfrac{1+\sum_{i=1}^{n} a_i x^i}{1+\sum_{i=1}^{n+1} b_i x^i}$.

Para $n = 2$, determine os valores de a_1, a_2, b_1, b_2 e b_3 da aproximação de Padé

$$e^x \approx \frac{1+a_1x+a_2x^2}{1+b_1x+b_2x^2+b_3x^3}$$

Problema 2.2 A expansão em série de potências da função erro $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_0^x e^{-t^2}dt$ é dada por $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}}\displaystyle\sum_{i=1}^{\infty} c_i x^i$, em que c_i é determinado recursivamente por $c_i = -\dfrac{2i-1}{i(2i+1)}c_{i-1}$, $c_0 = 1$ e $i = 1, 2, \cdots$. Determine os valores de a_0, a_1, a_2, b_1, b_1 e b_2 da aproximação de Padé

$$\operatorname{erf}(x) \approx \frac{2}{\sqrt{\pi}}\frac{a_0x+a_1x^3+a_2x^5}{1+b_1x^2+b_2x^4}$$

Problema 2.3 Fundamentado na seguinte expansão em série de potências do logaritmo neperiano de x: $\ln(x) \approx 2\sum_{k=1}^{n} \frac{1}{(2k-1)} \left(\frac{x-1}{x+1}\right)^{2k-1}$. Determine os valores de a_0, a_1, a_2, b_1, b_1 e b_2 da aproximação de Padé $\ln(x) \approx 2u\frac{a_0 + a_1u^2 + a_2u^4}{1 + b_1u^2 + b_2u^4}$ em que $u = \frac{x-1}{x+1}$. Analise e avalie o valor máximo (em módulo) do erro desta aproximação no intervalo $0,5 \leqslant x \leqslant 2$.

Problema 2.4 Considere as propriedades apresentadas abaixo que relacionam funções trigonométricas com as correspondentes funções hiperbólicas. Utilize as aproximações por frações continuadas apresentadas anteriormente para as funções trigonométricas tangente e arco tangente para obter as mesmas aproximações para as correspondentes funções hiperbólicas. Utilize também as aproximações por razão de polinômios apresentadas anteriormente para as funções trigonométricas cosseno, seno, tangente e arco tangente para obter as mesmas aproximações para as correspondentes funções hiperbólicas.

$$\begin{array}{ccc}
\cos(\boldsymbol{i}x) = \cosh(x) & \operatorname{sen}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{senh}(x) & \operatorname{tg}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{tgh}(x) \\
\cosh(\boldsymbol{i}x) = \cos(x) & \operatorname{senh}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{sen}(x) & \operatorname{tgh}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{tg}(x) \\
\operatorname{arcsen}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{arcsenh}(x) & \operatorname{arctg}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{arctgh}(x) & \operatorname{arctgh}(\boldsymbol{i}x) = \boldsymbol{i}\operatorname{arctg}(x) \\
\text{Sendo: } \boldsymbol{i} = \sqrt{-1} & \boldsymbol{i}^{2k} = (-1)^k & \boldsymbol{i}^{2k+1} = (-1)^k\boldsymbol{i}
\end{array}$$

Problema 2.5 A função integral exponencial de ordem n é definida por: $E_n(x) = \int_1^\infty \frac{e^{-tx}}{t^n} dt$.

A aproximação por frações continuadas desta função é dada por:

$$E_n(x) \approx \cfrac{e^{-x}}{x + \cfrac{n}{1 + \cfrac{1}{x + \cfrac{n+1}{1 + \cfrac{2}{x + \cfrac{n+2}{1 + \cfrac{3}{x + \cfrac{n+3}{1 + \cfrac{4}{x + \cfrac{n+4}{1 + \cdots}}}}}}}}}}$$

Obtenha a aproximação que permita que o módulo do erro seja inferior a 10^{-4} para $n = 1$ em todos os pontos tabelados abaixo.

x	0,0	0,5	1,0	1,5	2,0	2,5
$E_1(x)$	∞	0,5598	0,2194	0,1000	0,0489	0,02491

Problema 2.6 Obtenha a série de Fourier apropriada à função: $f(x) = \begin{cases} 0 \text{ para } -1 < x < 0 \\ x \text{ para } 0 < x < \frac{1}{2} \\ \frac{1}{2} \text{ para } \frac{1}{2} < x < 1 \end{cases}$.

Represente a série desenvolvida no intervalo $-4 < x < 4$.

3. Interpolação Polinomial

3.1 Introdução

O objetivo deste capítulo é buscar aproximações polinomiais de funções definidas e contínuas em um intervalo fechado $[a,b]$. A maior vantagem de empregar aproximações polinomiais de funções é a facilidade da determinação de suas derivadas e integrais que são também polinômios.

A busca de aproximações polinomiais é orientada pelo Teorema de Weierstrass[1].

Teorema 3.1.1 — Teorema de Weierstrass. Se $f(x)$ é uma função definida e contínua em um intervalo fechado $[a,b]$, então para cada $\varepsilon > 0$, existe um polinômio de grau $n(\varepsilon)$ tal que:

$$|f(x) - p_n(x)| < \varepsilon \ \ \forall x \in [a,b].$$

Apesar de o Teorema de Weierstrass estabelecer que existe um polinômio de grau $n(\varepsilon)$, a dificuldade é encontrar e estabelecer o grau de tal polinômio que satisfaça o critério relacionado ao erro da aproximação.

Dois tipos de problemas se apresentam na aproximação polinomial de funções em um intervalo fechado $[a,b]$. No primeiro a forma da função é definida e contínua em todo intervalo e o objetivo é a construção de um polinômio de grau n que satisfaça às $(n+1)$ condições $p_n(x_k) = f(x_k)$ para $k = 0, 1, 2, \cdots, n$, sendo os $(n+1)$ valores $\{x_0, x_1, \cdots, x_n\}$ chamados de **pontos nodais**. No segundo tipo de problema a forma da função $f(x)$ não é conhecida, sendo especificado apenas $(n+1)$ valores da mesma em $(n+1)$ pontos nodais distintos no intervalo fechado $[a,b]$. Neste tipo de problema o objetivo da determinação da aproximação polinomial é determinar o valor da função em pontos não tabelados, sendo por isto chamado de **polinômio interpolador**. Estes dois tipos de problemas são apresentados neste capítulo, assim como diferentes procedimentos adequados a cada problema.

Antes de se apresentar o real escopo deste capítulo dois comentários pertinentes são expostos a seguir.

1. Visando a generalização de alguns procedimentos numéricos é interessante torná-los independentes

[1] Karl Theodor Wilhelm Weierstrass (1815-1897).

dos valores das extremidades a e b do intervalo considerado, para isto dois procedimentos de *normalização* podem ser considerados:

(a) Transformação do intervalo $[a,b]$ para $[-1,+1]$, de acordo com $x = \frac{1-t}{2}a + \frac{1+t}{2}b = \frac{a+b}{2} + \frac{a-b}{2}t$ com $-1 \leq t \leq +1$, assim: $t = -1 \Rightarrow x = a$, $t = +1 \Rightarrow x = b$ e $t = \frac{2x-(a+b)}{b-a}$.

(b) Transformação do intervalo $[a,b]$ para $[0,+1]$, de acordo com $x = (1-t)a + tb = a + t(b-a)$ com $0 \leq t \leq 1$, assim: $t = 0 \Rightarrow x = a$, $t = 1 \Rightarrow x = b$ e $t = \frac{x-a}{b-a}$.

2. Um algoritmo bastante empregado em procedimentos numéricos para o cálculo de valores de polinômios é o **método de Horner**[2], que permite calcular o valor de polinômios de qualquer grau sem a necessidade de calcular a potência de seu argumento superior à primeira potência. Este procedimento é bastante adequado e acurado, minimizando possíveis erros de truncamento no cálculo. O método pode ser explicado através do cálculo do valor de um polinômio de grau n em um ponto $x = \alpha$, assim:
$p_n(\alpha) = \sum_{i=0}^{n} c_i \alpha^i = c_0 + c_1\alpha + c_2\alpha^2 + c_3\alpha^3 + \cdots + c_{n-1}\alpha^{n-1} + c_n\alpha^n$, que pode ser reagrupado na forma: $p_n(\alpha) = c_0 + \alpha(c_1 + \alpha(c_2 + \alpha(c_3 + \cdots + \alpha(c_{n-1} + \alpha c_n))\cdots)$.
Este procedimento pode ser programado de acordo com o algoritmo recursivo descrito abaixo.

$p \longleftarrow c_n$

Para $k = n-1, n-2, \cdots, 2, 1, 0,$ faça
$p \longleftarrow c_k + \alpha p$

3.2 Métodos Diretos de Determinação do Polinômio Interpolador

Uma vez que um polinômio de grau n contém $(n+1)$ coeficientes, para determiná-los são necessárias $(n+1)$ condições ou especificações, tal polinômio pode ser representado genericamente por:

$$p_n(x) = \sum_{k=0}^{n} c_k x^k \text{ sendo } c_k = \frac{1}{k!}\frac{d^k p_n(x)}{dx^k}\bigg|_{x=0}$$

Quando a aproximação polinomial é obtida através do polinômio de Taylor os coeficientes c_k são determinados por: $c_k = \frac{1}{k!}\frac{d^k f(x)}{dx^k}\Big|_{x=0}$ para $k = 0, 1, \cdots, n$ e o erro da aproximação é da forma: $R_n(x) = f(x) - p_n(x) = q(x)x^{n+1}$, o que implica em $\frac{d^k R_n(x)}{dx^k}\Big|_{x=0} = 0$ para $k = 0, 1, \cdots, n$.

Quando a aproximação polinomial é obtida através de uma *interpolação* polinomial de grau n os $(n+1)$ coeficientes c_k são determinados através das $(n+1)$ condições: $p_n(x_k) = f(x_k)$ para $k = 0, 1, \cdots, n$ e, em consequência, o erro da aproximação é da forma: $R_n(x) = f(x) - p_n(x) = q(x)\prod_{k=0}^{n}(x - x_k)$, o que implica em $R_n(x_k)$ para $k = 0, 1, \cdots, n$. O polinômio de grau $(n+1)$ da expressão do erro da aproximação $\prod_{k=0}^{n}(x - x_k)$ é chamado de **polinômio nodal**, $p_{nodal}(x)$, cujas raízes são os pontos nodais adotados.

No Exemplo 3.1, aproximações polinomiais de terceiro grau de uma função específica feitas pelo polinômio de Taylor e por interpolação polinomial são confrontadas.

[2] William George Horner (1786-1837).

■ **Exemplo 3.1** Aproximação polinomial de terceiro grau da função $f(x) = \dfrac{20x}{1+20x^2}$ no intervalo fechado $[0,1]$.

- Polinômio de Taylor de terceiro grau: $p_3(x) = 20x - 400x^3$.
- Polinômio interpolador de terceiro grau adotando quatro pontos igualmente espaçados no intervalo (incluindo as duas extremidades) $x_0 = 0, x_1 = \frac{1}{3}, x_2 = \frac{2}{3}$ e $x_3 = 1$. Os coeficientes são determinados pela resolução do sistema linear

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{1}{3} & \frac{1}{9} & \frac{1}{27} \\ 1 & \frac{2}{3} & \frac{4}{9} & \frac{8}{27} \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} f(0) \\ f(\frac{1}{3}) \\ f(\frac{2}{3}) \\ f(1) \end{bmatrix} \Rightarrow \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 13,506 \\ -26,568 \\ 14,015 \end{bmatrix}$$

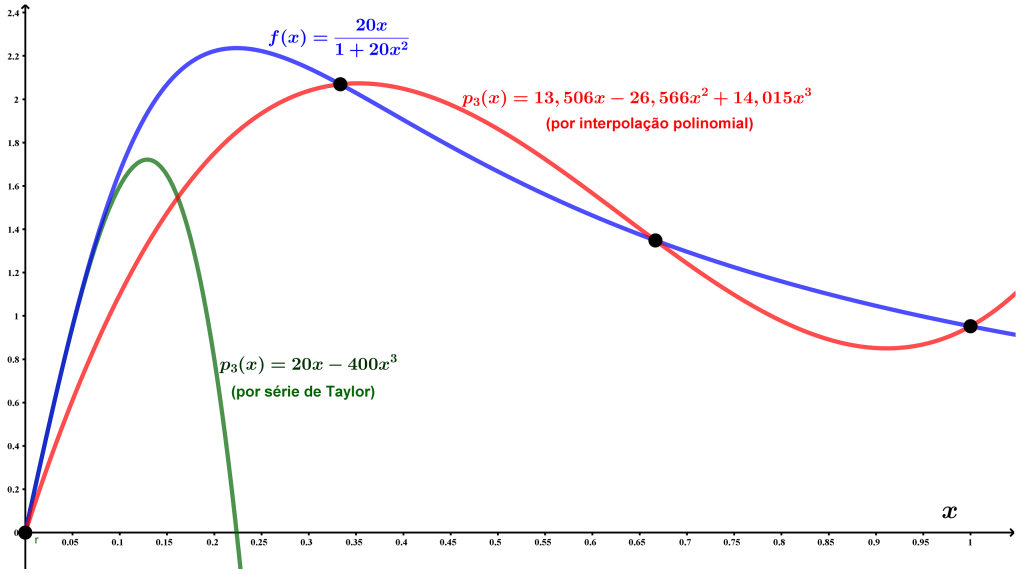


Figura 3.1: Aproximações polinomiais de terceiro grau de $f(x) = \dfrac{20x}{1+20x^2}$.

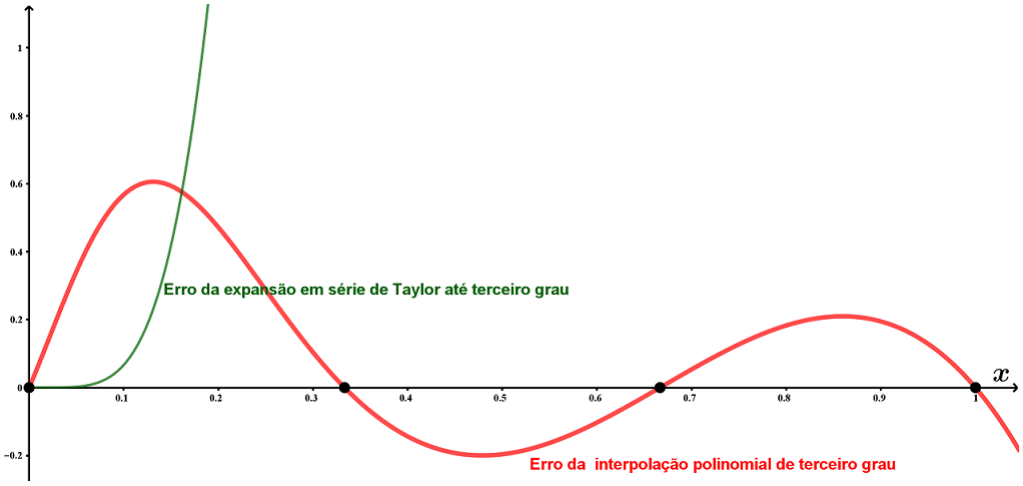


Figura 3.2: Erros de aproximações polinomiais de terceiro grau de $f(x) = \dfrac{20x}{1+20x^2}$.

As Figuras 3.1 e 3.2 mostram claramente que a aproximação polinomial da função pelo

polinômio de Taylor só apresenta valores pequenos do erro em pontos próximos ao ponto $x = 0$, enquanto que a aproximação pelo polinômio interpolador de terceiro grau apresenta valores razoáveis do erro em todo o intervalo. ∎

Na determinação direta do polinômio de grau n que satisfaz às $(n+1)$ condições $p_n(x_k) = f(x_k)$ para $k = 0, 1, 2, \cdots, n$, os coeficientes do polinômio são calculados pela resolução do sistema linear de equações:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} . \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}$$

Ou $\mathbf{A} \cdot \mathbf{c} = \mathbf{y}$ sendo $A_{i,j} = x_i^j$ e $y_i = f(x_i)$ para $i, j = 0, 1, 2, \cdots, n$.

A matriz $\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}$ é chamada de **matriz de Vandermonde**[3], matriz em que os termos de cada linha estão em progressão geométrica, seu determinante é expresso por:

$$|\mathbf{A}| = \prod_{0 \le i < j \le n} (x_j - x_i).$$

Quando se utiliza um número elevado de pontos nodais ou pontos nodais muito próximos tem-se um baixo valor numérico do determinante da matriz de Vandermonde, o que é um indicativo do mau condicionamento da matriz, tornando inapropriado o emprego de métodos de eliminação de Gauss sem pivotamento para a resolução do correspondente sistema linear (ver Seção 5.3).

■ **Exemplo 3.2** A solução analítica do problema de reação-difusão em um partícula catalítica esférica isotérmica com reação de primeira ordem é dada pela seguinte expressão:

$$f(x) = \frac{\operatorname{senh}(\Phi x)}{x \operatorname{senh}(\Phi)}$$

em que x é o raio adimensional da partícula, $f(x)$ é a concentração adimensional do reagente e $\Phi = 5$ é o módulo de Thiele[4]. Utilizando como pontos nodais, pontos igualmente espaçados entre 0,1 e 0,9, com espaçamento uniforme de 0,1 para o caso (a) e de 0,04 para o caso (b), foram obtidos os polinômios interpoladores de graus 8 e 20, respectivamente. Após obter os polinômios, os mesmos foram utilizados para obter os valores aproximados no intervalo de 0 a 1 em espaçamento uniforme de 0,01. Note que entre 0 e 0,1 e entre 0,9 e 1 os valores são extrapolados. Nas Figuras 3.3 e 3.4 é mostrado que o método direto falha na determinação do polinômio interpolador de grau 20 devido ao mau condicionamento da matriz de Vandermonde ($\kappa = 2,8 \times 10^{18}$, ver Seção 5.2). Porém, ao resolver o sistema linear com a técnica de pivotamento (ver Seção 5.3), foi possível sanar o problema de mau condicionamento, conforme pode ser visto nessas figuras. Observa-se também na Figura 3.4 o crescimento do erro da aproximação do caso (a) nos intervalos de extrapolação.

∎

[3] Alexandre-Théophile Vandermonde (1735-1796).

[4] Ernest William Thiele (1895-1993).

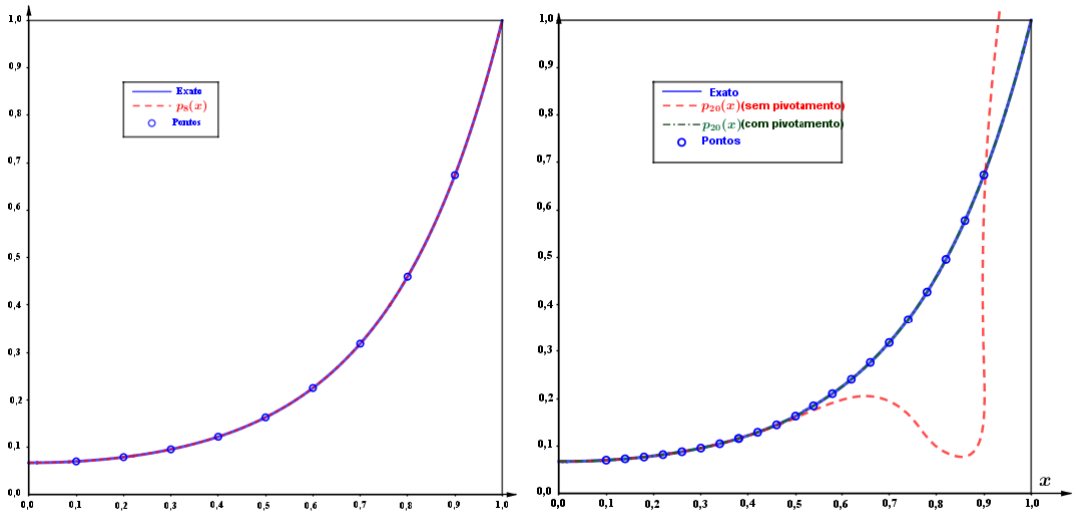


Figura 3.3: Aproximações polinomiais da função do Exemplo 3.2.

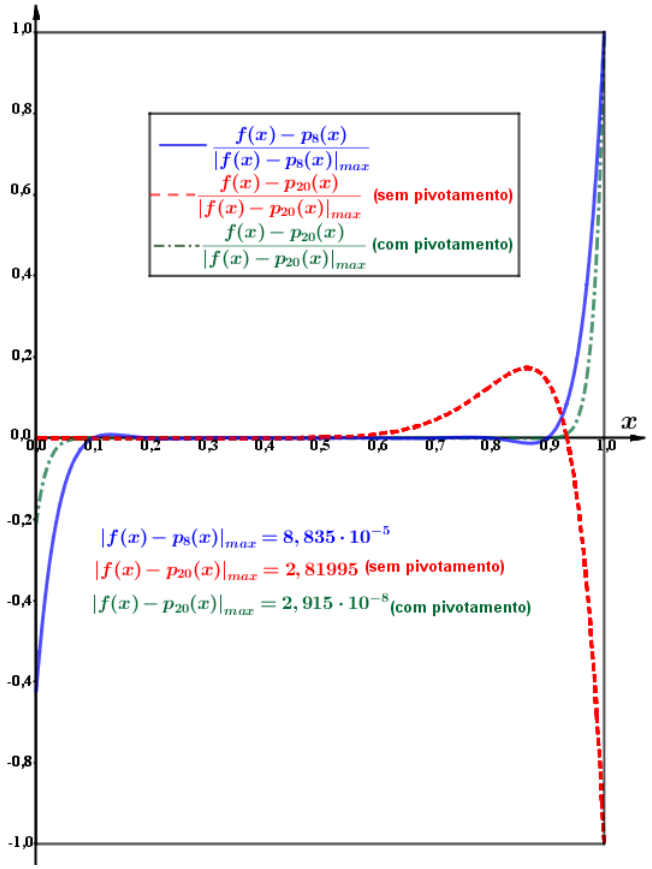


Figura 3.4: Erros de aproximações polinomiais da função do Exemplo 3.2.

3.3 Tabela de Diferenças Divididas de Newton

O **Método das Diferenças Divididas de Newton** consiste em *construir* um polinômio interpolador de grau n de uma função definida e contínua $f(x)$ em um intervalo fechado $[a,b]$, fundamentado

em $(n+1)$ valores da função em pontos nodais $a \leq x_k \leq b\,[k = 0, 1, 2, \cdots, n]$ expresso na forma:

$$p_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_n(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})$$

Com os valores dos coeficientes a_i conhecidos, o cômputo do valor do polinômio de Newton de grau n para $x = \alpha$ pode ser feito por um procedimento recursivo semelhante ao método de Horner pelo algoritmo:

$p_n \longleftarrow a_n$

Para $k = n-1, n-2, \cdots, 2, 1, 0,$ faça
$p_n \longleftarrow a_k + (\alpha - x_k)p_n$

Tendo em vista que $p_n(x_k) = f(x_k)$, $k = 0, 1, 2, \cdots, n$, resulta em um sistema linear triangular inferior:

$$\begin{aligned}
&a_0 = f(x_0)\\
&a_0 + (x_1-x_0)a_1 = f(x_1)\\
&a_0 + (x_2-x_0)a_1 + (x_2-x_0)(x_2-x_1)a_2 = f(x_2)\\
&a_0 + (x_3-x_0)a_1 + (x_3-x_0)(x_3-x_1)a_2 + (x_3-x_0)(x_3-x_1)(x_3-x_2)a_3 = f(x_3)\\
&\vdots\\
&a_0 + (x_n-x_0)a_1 + (x_n-x_0)(x_n-x_1)a_2 + (x_n-x_0)(x_n-x_1)(x_n-x_2)a_3 + \cdots\\
&+(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})a_n = f(x_n)
\end{aligned}$$

Este sistema pode ser resolvido na forma recursiva:

$$\begin{aligned}
a_0 &= f(x_0)\\
a_1 &= \frac{f(x_1)-a_0}{(x_1-x_0)}\\
a_2 &= \frac{f(x_2)-a_0-(x_2-x_0)a_1}{(x_2-x_0)(x_2-x_1)}\\
a_3 &= \frac{f(x_3)-a_0-(x_3-x_0)a_1-(x_3-x_0)(x_3-x_1)a_2}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\
&\vdots\\
a_n &= \frac{f(x_n)-a_0-(x_n-x_0)a_1-(x_n-x_0)(x_n-x_1)a_2-(x_n-x_0)(x_n-x_1)(x_n-x_2)a_3-\cdots}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}
\end{aligned}$$

Definindo as funções diferenças divididas:

$$\begin{aligned}
f[x_k,x_{k-1}] &= \frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}\,k = 1, 2, \cdots, n\\
f[x_k,x_{k-1},x_{k-2}] &= \frac{f[x_k,x_{k-1}]-f[x_{k-1},x_{k-2}]}{x_k-x_{k-2}}\,k = 2, 3, \cdots, n\\
f[x_k,x_{k-1},x_{k-2},x_{k-3}] &= \frac{f[x_k,x_{k-1},x_{k-2}]-f[x_{k-1},x_{k-2},x_{k-3}]}{x_k-x_{k-3}}\,k = 3, 4, \cdots, n\\
&\vdots\\
f[x_n,x_{n-1},\cdots,x_1,x_0] &= \frac{f[x_n,x_{n-1},\cdots,x_1]-f[x_{n-1},\cdots,x_1,x_0]}{x_n-x_0}
\end{aligned}$$

Temos então:

$$\begin{aligned}
a_0 &= f(x_0)\\
a_1 &= f[x_1,x_0]\\
a_2 &= f[x_2,x_1,x_0]\\
a_3 &= f[x_3,x_2,x_1,x_0]\\
&\vdots\\
a_n &= f[x_n,x_{n-1},\cdots,x_1,x_0]
\end{aligned}$$

Na Figura 3.5 é mostrada a estrutura de uma **Tabela de Diferenças Divididas de Newton**.

x_k	$f(x_k)$	$f[x_k,x_{k-1}]$	$f[x_k,x_{k-1},x_{k-2}]$	$f[x_k,x_{k-1},x_{k-2},x_{k-3}]$	$f[x_k,x_{k-1},x_{k-2},x_{k-3},x_{k-4}]$
x_0	$f(x_0)$				
		$f[x_1,x_0]$			
x_1	$f(x_1)$		$f[x_2,x_1,x_0]$		
		$f[x_2,x_1]$		$f[x_3,x_2,x_1,x_0]$	
x_2	$f(x_2)$		$f[x_3,x_2,x_1]$		$f[x_4,x_3,x_2,x_1,x_0]$
		$f[x_3,x_2]$		$f[x_4,x_3,x_2,x_1]$	
x_3	$f(x_3)$		$f[x_4,x_3,x_2]$		
		$f[x_4,x_3]$			
x_4	$f(x_4)$				

Figura 3.5: Estrutura de uma Tabela de Diferenças Divididas de Newton.

Denomina-se **diferença dividida de ordem *k*** a diferença dividida com ***k*+1** argumentos, assim a coluna da tabela $f[x_k,x_{k-1}]$ é a coluna com as diferenças divididas de primeira ordem, a coluna da tabela $f[x_k,x_{k-1},x_{k-2}]$ é a coluna com as diferenças divididas de segunda ordem, a coluna da tabela $f[x_k,x_{k-1},x_{k-2},x_{k-3}]$ é a coluna com as diferenças divididas de terceira ordem e assim por diante.

É importante ressaltar que as diferenças divididas são versões *discretas* das derivadas contínuas, assim:

Se $f(x)=x$ então $f[a,b]=\dfrac{a-b}{a-b}=1$

Se $f(x)=x^2$ então $f[a,b]=\dfrac{a^2-b^2}{a-b}=a+b \Rightarrow f[a,b,c]=\dfrac{a+b-(b+c)}{a-c}=1$

Se $f(x)=x^3$ então $f[a,b]=\dfrac{a^3-b^3}{a-b}=a^2+b^2+ab \Rightarrow$

$$f[a,b,c]=\frac{a^2+b^2+ab-(b^2+c^2+bc)}{a-c}=a+b+c \Rightarrow f[a,b,c,d]=\frac{a+b+c-(b+c+d)}{a-d}=1$$

Deste modo, se a coluna de diferenças divididas de ordem n na tabela de diferenças divididas for praticamente constante é um indicativo que a função $f(x)$ é um polinômio de grau n.

■ **Exemplo 3.3** Obtenção do polinômio interpolador de grau 3 usando as fórmulas das diferenças divididas de Newton para os dados tabelados a seguir.

k	x_k	$f(x_k)$	Δ_1	Δ_2	Δ_3
0	0	-5			
			6		
1	1	1		2	
			12		1
2	3	25		6	
			30		
3	4	55			

Tomando como base o ponto x_0: $p_3(x) =$
$f(x_0)+f[x_1,x_0](x-x_0)+f[x_2,x_1,x_0](x-x_1)(x-x_0)+f[x_3,x_2,x_1,x_0](x-x_2)(x-x_1)(x-x_0)$
$p_3(x) = -5+6(x-0)+2(x-1)(x-0)+1(x-3)(x-1)(x-0) = x^3-2x^2+7x-5$

Tomando como base o ponto x_3: $p_3(x) =$
$f(x_3)+f[x_3,x_2](x-x_3)+f[x_3,x_2,x_1](x-x_3)(x-x_2)+f[x_3,x_2,x_1,x_0](x-x_3)(x-x_2)(x-x_1)$
$p_3(x) = 55+30(x-4)+6(x-4)(x-3)+1(x-4)(x-3)(x-1) = x^3-2x^2+7x-5$

■

Uma forma recursiva de determinação dos coeficientes da aproximação polinomial de Newton pode ser implementada e entendida considerando as sucessivas interpolações fundamentadas em $(n+1)$ pontos nodais $\{x_0,x_1,x_2,\cdots,x_n\}$.

- Interpolação de ordem zero $p_0(x) = f(x_0) = f[x_0] \Rightarrow a_0 = p_0(x_0) = f(x_0)$
- Interpolação linear $p_1(x) = p_0(x)+f[x_1,x_0](x-x_0)$, condições a serem satisfeitas:
  $\begin{cases} p_1(x_0) = f(x_0) \\ p_1(x_1) = f(x_1) \end{cases}$, a primeira condição já é satisfeita e para que: $p_1(x_1) = f(x_1)$
  deve-se ter: $f(x_1) = p_0(x_1)+f[x_1,x_0](x_1-x_0) \Rightarrow a_1 = f[x_1,x_0] = \dfrac{f(x_1)-p_0(x_1)}{x_1-x_0}$
- Interpolação parabólica $p_2(x) = p_1(x)+f[x_2,x_1,x_0](x-x_0)(x-x_1)$, condições a serem satisfeitas: $\begin{cases} p_2(x_0) = f(x_0) \\ p_2(x_1) = f(x_1) \\ p_2(x_2) = f(x_2) \end{cases}$, as duas primeiras condições já são satisfeitas e para que: $p_2(x_2) = f(x_2)$, deve-se ter: $a_2 = f[x_2,x_1,x_0] = \dfrac{f(x_2)-p_1(x_2)}{(x_2-x_0)(x_2-x_1)}$.
- Interpolação cúbica $p_3(x) = p_2(x)+f[x_3,x_2,x_1,x_0](x-x_0)(x-x_1)(x-x_2)$, condições a serem satisfeitas: $\begin{cases} p_3(x_0) = f(x_0) \\ p_3(x_1) = f(x_1) \\ p_3(x_2) = f(x_2) \\ p_3(x_3) = f(x_3) \end{cases}$ as três primeiras condições já são satisfeitas e para que: $p_3(x_3) = f(x_3)$, deve-se ter: $a_3 = f[x_3,x_2,x_1,x_0] = \dfrac{f(x_3)-p_2(x_3)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$.
- Interpolação de grau n
  $p_n(x) = p_{n-1}(x)+f[x_n,x_{n-1},\cdots,x_2,x_1,x_0](x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})$,
  condições a serem satisfeitas: $\begin{cases} p_n(x_0) = f(x_0) \\ p_n(x_1) = f(x_1) \\ p_n(x_2) = f(x_2) \\ \vdots \\ p_n(x_n) = f(x_n) \end{cases}$ as $(n-1)$ primeiras condições já são satisfeitas e para que: $p_n(x_n) = f(x_n)$, deve-se ter:

$$a_n = f[x_n,x_{n-1},\cdots,x_2,x_1,x_0] = \frac{f(x_n)-p_{n-1}(x_n)}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}.$$

Permitindo sugerir a seguinte forma algorítmica:

$a_0 \longleftarrow f(x_0)$

Para $i = 1,\cdots,n,$ faça

$$a_i \longleftarrow \frac{f(x_i)-p_{i-1}(x_i)}{\prod_{k=0}^{i-1}(x_i-x_k)}$$

Note que antes de se implementar este algoritmo, o procedimento de cálculo do polinômio de Newton de grau n genérico deve já ter sido implementado.

Os coeficientes $a_0, a_1, a_2, \cdots, a_n$ do polinômio podem também ser calculados diretamente (sem a construção da tabela de diferenças divididas) pela resolução do sistema linear triangular inferior pelo procedimento recursivo:

$$a_0 = f(x_0)$$
$$a_k = \frac{f(x_k) - \sum_{i=0}^{k-1} A_{k,i} a_i}{A_{k,k}} \quad \text{para } k = 1, 2, \cdots, n$$

A matriz triangular inferior, **A**, do sistema depende apenas dos valores dos pontos nodais e é determinada pelo seguinte algoritmo:

> Para $j = 1, \cdots, n,$ faça
> $A_{j,0} \longleftarrow 1$

> Para $j = 1, \cdots, n,$ faça
> para $i = j, \cdots, n,$ faça
> $A_{i,j} \longleftarrow A_{i,j-1}(x_i - x_{j-1})$

3.4 Interpolação Polinomial de Lagrange

Além do método de determinação do polinômio interpolador pelas diferenças divididas, outra forma que também evita o procedimento de resolução de um sistema linear é o do emprego de uma base de expansão composta pelos **polinômios interpoladores de Lagrange**. Quando se utilizam $(n+1)$ pontos nodais $(x_0, x_1, x_2, \cdots, x_n)$, tais polinômios são polinômios de grau n que apresentam a propriedade:

$$\ell_j(x) = \begin{cases} 1 \text{ para } x = x_j \\ 0 \text{ para } x = x_i \neq x_j \end{cases} \qquad \text{para } i, j = 0, 1, 2, \cdots, n$$

ou, em notação mais compacta: $\ell_j(x_i) = \delta_{i,j}$.

Os polinômios de Lagrange são fatorados na forma: $\ell_j(x) = \prod_{k=0 \neq j}^{n} \frac{x - x_k}{x_j - x_k}$ para $j = 0, 1, 2, \cdots, n$.

O polinômio interpolador pode então ser expresso por: $p_n(x) = \sum_{j=0}^{n} \ell_j(x) f(x_j)$.

■ **Exemplo 3.4** Obtenção do polinômio interpolador de Lagrange de grau 2 para os seguintes dados:

k	x_k	$f(x_k)$
0	0	-5
1	1	1
2	3	25

$$\ell_0 = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{x^2-4x+3}{3}$$
$$\ell_1 = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-0)(x-3)}{(1-0)(1-3)} = \frac{-x^2+3x}{2}$$
$$\ell_1 = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{x^2-x}{6}$$

$$p_2(x) = \sum_{j=0}^{2} \ell_j(x) f(x_j) = -5\ell_0(x) + \ell_1(x) + 25\ell_2(x) = 2x^2 + 4x - 5 \qquad \blacksquare$$

Definindo o *polinômio nodal*: $p_{nodal}(x) = \prod_{k=0}^{n}(x - x_k)$, polinômio da grau $(n+1)$ cujas raízes são todos os pontos nodais, pode-se expressar os polinômios de Lagrange na forma:

$$\ell_j(x) = \frac{p_{nodal}(x)}{(x - x_j)p'_{nodal}(x_j)} \text{ para } j = 0, 1, 2, \cdots, n.$$

O algoritmo da interpolação polinomial de Lagrange é a seguir descrito.

Dado $(n+1)$ pontos $\{x_i, y_i\}$, para o cálculo do valor interpolado para $x = \alpha$ assim se procede:

> Para $i = 0, 1, \cdots, n,$ faça
> $p_i \longleftarrow 1$
> para $j = 1, \cdots, n,$ faça
> $$p_i \longleftarrow \left(\frac{\alpha - x_j}{x_i - x_j}\right) p_i \quad \text{se } j \neq i$$

$Y \longleftarrow 0$

> Para $i = 0, 1, \cdots, n,,$ faça
> $Y \longleftarrow Y + p_i y_i$

Ao final do algoritmo Y contém o valor interpolado de $f(x)$ em $x = \alpha$.

Uma forma iterativa de gerar os polinômios de Lagrange foi proposta por Neville[5], que além de *algoritmizar* a geração dos polinômios de Lagrange permite estimar a acurácia da interpolação e selecionar os pontos nodais necessários à acurácia desejada (Bulirsch e Stoer, 1966; Weisstein, 2006).

Teorema 3.4.1 — Teorema de Neville. Se $p_n(x)$ é o polinômio interpolador de grau n de uma função $f(x)$ gerado pelos valores da função em $(n+1)$ pontos nodais distintos $\{x_0, x_1, x_2, \cdots, x_n\}$ nos quais $\{y_0 = f(x_0), y_1 = f(x_1), y_2 = f(x_2), \cdots, y_n = f(x_n)\}$, então

$$p_n(x) = \frac{(x - x_j)p_{012\cdots(j-1)(j+1)\cdots n} - (x - x_i)p_{012\cdots(i-1)(i+1)\cdots n}}{(x_i - x_j)}$$

Para exemplificar o procedimento a função do Exemplo 3.1 é considerada: $f(x) = \dfrac{20x}{1 + 20x^2}$ e os pontos nodais: $x = \begin{bmatrix} 0,0 & 0,2 & 0,4 & 0,6 & 0,8 & 1,0 \end{bmatrix}$.

Tem-se:

$$p_{0123} = \frac{(x - 0,2)(x - 0,4)(x - 0,6)}{0,2.0,4.0,6} f(0) + \frac{x(x - 0,4)(x - 0,6)}{(-0,2).0,2.0,4} f(0,2) +$$

$$+ \frac{x(x - 0,2)(x - 0,6)}{(-0,4).(-0,2).0,2} f(0,4) + \frac{x(x - 0,2)(x - 0,4)}{(-0,6).(-0,4).(-0,2)} f(0,6)$$

Para simplificar a notação considera-se a matriz $Q_{i,j} = p_{i-j,i-j+1,\cdots,i-1,i}$ para $i = 0, 1, \cdots, n$ e $j \leq i \Rightarrow Q_{i,0} = p_0 = f(x_0)$.

Aplicando o Teorema de Neville a esta notação: $Q_{i,j} = \dfrac{(x - x_{i-j})Q_{i,j-1} - x_i Q_{i-1,j-1}}{x_i - x_{i-j}}$.

[5] Eric Harold Neville (1889–1961).

Baseado nesta última equação, configura-se o seguinte procedimento recursivo para a geração da matriz **Q**.

> Para $i = 0, 1, \cdots, n,$ faça
> $Q_{i,0} \longleftarrow y_i$
> para $j = 1, \cdots, n,$ faça
> $Q_{i,j} \longleftarrow 0$

> Para $i = 0, 1, \cdots, n,,$ faça
> para $j = 1, \cdots, i,$ faça
> $$Q_{i,j} \longleftarrow \frac{(x - x_{i-j})Q_{i,j-1} - x_i Q_{i-1,j-1}}{x_i - x_{i-j}}$$

O procedimento é aplicado à função do Exemplo 3.1 $f(x) = \dfrac{20x}{1+20x^2}$, os pontos nodais: $x = \begin{bmatrix} 0,0 & 0,2 & 0,4 & 0,6 & 0,8 & 1,0 \end{bmatrix}$ e $x = 0,5$, dando origem à matriz **Q**:

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 2,222222 & 5,555556 & 0 & 0 & 0 & 0 \\ 1,904762 & 1,746032 & 0,793651 & 0 & 0 & 0 \\ 1,463415 & 1,684088 & 1,699574 & 1,548587 & 0 & 0 \\ 1,15942 & 1,615412 & 1,666919 & 1,683247 & 1,632749 & 0 \\ 0,952381 & 1,469979 & 1,65177 & 1,664394 & 1,676177 & 1.654463 \end{bmatrix}$$

O valor verdadeiro de $f(0,5)$ é igual a $\frac{5}{3} = 1,666667$. A quarta linha da segunda coluna da matriz **Q** é o valor da interpolação linear de $f(x)$ utilizando os pontos: $\{0,4\ 0,6\}$ que é o único par de pontos que envolve o valor 0,5; a terceira coluna lista os valores das interpolações quadráticas usando três pontos nodais, há apenas dois trios que envolvem o ponto 0,5 que são $\{0,2\ 0,4\ 0,6\}$ e $\{0,4\ 0,6\ 0,8\}$ e os correspondentes valores interpolados localizam-se na quarta e quinta linha desta coluna; a penúltima coluna apresenta os valores das interpolações cúbicas usando quatro pontos que sempre envolvem o valor 0,5 e, finalmente, o valor listado na última coluna representa o valor do polinômio interpolador de quarto grau empregando todos os pontos nodais.

Os valores das derivadas da interpolação em cada ponto nodal pode ser calculada por:

$$\left.\frac{dp_n(x)}{dx}\right|_{x_i} = \sum_{j=0}^{n} \left.\frac{d\ell_j(x)}{dx}\right|_{x_i} f(x_j) = \sum_{j=0}^{n} A_{i,j} f(x_j) \text{ sendo } A_{i,j} = \left.\frac{d\ell_j(x)}{dx}\right|_{x_i}.$$

Em vista de: $\ell_j(x) = \dfrac{p_{nodal}(x)}{(x-x_j)p'_{nodal}(x_j)} \Rightarrow (x-x_j)\ell_j(x) = \dfrac{p_{nodal}(x)}{p'_{nodal}(x_j)}$ logo:

$$(x-x_j)\frac{d\ell_j(x)}{dx} + \ell_j(x) = \frac{p'_{nodal}(x)}{p'_{nodal}(x_j)} \Rightarrow A_{i,j} = \frac{p'_{nodal}(x_i)}{(x_i-x_j)p'_{nodal}(x_j)} \text{ para } i \neq j$$

Para o cálculo dos elementos da diagonal da matriz **A**, utiliza-se a propriedade dos polinômio interpoladores de Lagrange: $1 = \sum_{j=0}^{n} \ell_j(x) \Rightarrow \sum_{j=0}^{n} \dfrac{d\ell_j(x)}{dx} = 0, \forall x$ no caso particular de $x = x_i$, tem-se:

$$\sum_{j=0}^{n} A_{i,j} = 0 \Rightarrow A_{i,i} = -\sum_{j=0 \neq i}^{n} A_{i,j}.$$

Obtendo-se: $A_{i,j} = \begin{cases} \dfrac{p'_{nodal}(x_i)}{(x_i-x_j)p'_{nodal}(x_j)} \text{ para } i \neq j \\ -\displaystyle\sum_{j=0 \neq i}^{n} A_{i,j} \text{ para } i = j \end{cases}$ para $i, j = 0, 1, 2, \cdots, n.$

O pseudo-código de geração da matriz **A** é a seguir descrito.

Para $i = 0, 1, \cdots, n,$ faça
$P \longleftarrow 1$
$c_i \longleftarrow 0$
para $j = 0, \cdots, n,$ faça
$p \longleftarrow (x_i - x_j)$
$c_i \longleftarrow P + p\, c_i$
$P \longleftarrow p\, P$

Para $i = 0, 1, \cdots, n,,$ faça
$A_{i,i} \longleftarrow 0$
para $j = 0, 1, \cdots, n,$ faça
$A_{i,j} \longleftarrow \dfrac{c_i}{(x_i - x_j)c_j}$ se $j \neq i$
$A_{i,i} \longleftarrow A_{i,i} - A_{i,j}$

Além da geração da matriz **A**, este procedimento determina os valores das derivadas do polinômio nodal em cada ponto nodal e estes valores encontram-se armazenados no vetor **c**, isto é $c_i = p'_{nodal}(x_i)$ para $i = 0, 1, \cdots, n$.

Definindo os vetores: $\mathbf{y}_i = f(x_i)$ e $\left.\dfrac{d\mathbf{y}}{dx}\right|_i = \left.\dfrac{dp_n(x)}{dx}\right|_{x_i}$, resulta em

$\dfrac{d\mathbf{y}}{dx} = \mathbf{A}\,\mathbf{y}$ e $\dfrac{d^k\mathbf{y}}{dx^k} = \mathbf{A}^k\,\mathbf{y}$ para $k = 1, 2, \cdots, n$.

Definindo a matriz **G** cuja coluna k é : $\mathbf{G}^{<k>} = \dfrac{1}{k!}\mathbf{A}^k\,\mathbf{y} = \dfrac{d^k\mathbf{y}}{dx^k}$ obtém-se, em cada linha i, os valores dos coeficientes da expansão de $p_n(x)$ em torno do ponto nodal x_i, isto é:

$$p_n(x) = \sum_{j=0}^{n} G_{i,j}(x - x_i)^j \quad \text{para } i = 0, 1, 2, \cdots, n.$$

A matriz **G** é determinada pelo procedimento recursivo: $\mathbf{G}^{<k>} = \dfrac{1}{k}\mathbf{A}\,\mathbf{G}^{<k-1>}$ para $k = 1, 2, \cdots, n,$ iniciando com $\mathbf{G}^{<0>} = \mathbf{y}$.

O procedimento é aplicado à função do Exemplo 3.1 $f(x) = \dfrac{20x}{1 + 20x^2}$ com os pontos nodais: $\begin{bmatrix} 0{,}0 & 0{,}2 & 0{,}4 & 0{,}6 & 0{,}8 & 1{,}0 \end{bmatrix}$, dando origem à matriz **G**:

$$\mathbf{G} = \begin{bmatrix} 0 & 26{,}032756 & -105{,}932322 & 185{,}213239 & -152{,}614363 & 48{,}253071 \\ 2{,}222222 & 1{,}387781 & -27{,}57158 & 82{,}422977 & -104{,}361292 & 48{,}253071 \\ 1{,}904762 & -2{,}703631 & 0{,}695742 & 18{,}235172 & -56{,}108222 & 48{,}253071 \\ 1{,}463415 & -1{,}646552 & 2{,}031118 & -7{,}350177 & -7{,}855151 & 48{,}253071 \\ 1{,}15942 & -1{,}581466 & -0{,}403979 & 5{,}66693 & 40{,}39792 & 48{,}253071 \\ 0{,}952381 & 0{,}615732 & 16{,}551925 & 57{,}286494 & 88{,}65099 & 48{,}253071 \end{bmatrix}.$$

Valores próximos de zero de **todos** os elementos das últimas colunas da matriz **G** é um indicativo que a função é um polinômio de grau $(m-2)$, sendo m a posição da primeira coluna próxima de zero. Para ilustrar este comportamento calcula-se a matriz **G** para a função $f(x) = x^3 - 2x^2 + 3x + 10$ com os pontos nodais $\begin{bmatrix} 0{,}0 & 0{,}2 & 0{,}4 & 0{,}6 & 0{,}8 & 1{,}0 \end{bmatrix}$,

$$\mathbf{G} = \begin{bmatrix} 10 & 3 & -2 & 1 & 0 & 0 \\ 10{,}528 & 2{,}32 & -1{,}4 & 1 & 0 & 0 \\ 10{,}944 & 1{,}88 & -0{,}8 & 1 & 0 & 0 \\ 11{,}296 & 1{,}68 & -0{,}2 & 1 & 0 & 0 \\ 11{,}632 & 1{,}72 & 0{,}4 & 1 & 0 & 0 \\ 12 & 2 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

Outra forma de interpolação polinomial semelhante à interpolação de Lagrange é a **Interpolação de Hermite**[6]. A diferença dos dois procedimentos reside no fato de o polinômio interpolador de Hermite ser de grau $(2n+1)$, que utiliza os valores da função e de sua derivada em $(n+1)$ pontos nodais distintos $\{x_0, x_1, x_2, \cdots, x_n\}$. Assim: $p_{2n+1}(x_k) = f(x_k)$ e $p'_{2n+1}(x_k) = f'(x_k)$ para $k = 0, 1, 2, \cdots, n$. De forma análoga à do polinômio interpolador de Lagrange, O polinômio interpolador de Hermite pode ser expresso na forma:

$$p_{2n+1}(x) = \sum_{j=0}^{n} q_j(x) f(x_j) + \sum_{j=0}^{n} r_j(x) f'(x_j).$$

$$\begin{cases} q_j(x_i) = \delta_{ij} \\ q'_j(x_i) = 0 \end{cases} \qquad \begin{cases} r_j(x_i) = 0 \\ r'_j(x_i) = \delta_{ij} \end{cases}$$

Estas condições permitem propor as formas: $\begin{cases} q_j(x) = [1+\alpha_j(x-x_j)]\ell_j(x)^2 \\ r_j(x) = \beta_j(x-x_j)\ell_j(x)^2 \end{cases}$.

Em vista de $q'_j(x_j) = 0$, obtém-se:

$q'_j(x_j) = \left\{2[1+\alpha_j(x-x_j)]\ell'_j(x) + \alpha_j\ell_j(x)\right\}_{x=x_j} \ell_j(x_j) = 2\ell'_j(x_j) + \alpha_j = 0,$

logo $\alpha_j = -2\ell'_j(x_j) = -2A_{jj}$

Em vista de $r'_j(x_j) = 1$, obtém-se:

$r'_j(x_j) = \beta_j \left\{2(x-x_j)\ell'_j(x) + \ell_j(x)\right\}_{x=x_j} \ell_j(x_j) = \beta_j = 1.$

Resultando em: $\begin{cases} q_j(x) = [1-2A_{jj}(x-x_j)]\ell_j(x)^2 \\ r_j(x) = (x-x_j)\ell_j(x)^2 \end{cases}$, e

$$p_{2n+1}(x) = \sum_{j=0}^{n} [1-2A_{jj}(x-x_j)]\ell_j(x)^2 f(x_j) + \sum_{j=0}^{n} (x-x_j)\ell_j(x)^2 f'(x_j).$$

Em vista de: $\ell_j(x) = \dfrac{p_{nodal}(x)}{(x-x_j)p'_{nodal}(x_j)} \Rightarrow (x-x_j)\ell_j(x) = \dfrac{p_{nodal}(x)}{p'_{nodal}(x_j)}$ permitindo expressar:

$$\begin{cases} q_j(x) = [1-2A_{jj}(x-x_j)]\ell_j(x)^2 = \ell_j(x)^2 - 2\dfrac{A_{jj}\ell_j(x)}{p'_{nodal}(x_j)} p_{nodal}(x) \\ r_j(x) = (x-x_j)\ell_j(x)^2 = \dfrac{\ell_j(x)}{p'_{nodal}(x_j)} p_{nodal}(x) \end{cases} \quad \text{e}$$

$$p_{2n+1}(x) = \sum_{j=0}^{n} \ell_j(x)^2 f(x_j) + \left[\sum_{j=0}^{n} \frac{f'(x_j) - 2A_{j,j}f(x_j)}{p'_{nodal}(x_j)} \ell_j(x)\right] p_{nodal}(x).$$

3.5 Análise dos Erros da Interpolação Polinomial

Nas interpolações polinomiais de grau n usando $(n+1)$ pontos nodais distintos $\{x_0, x_1, x_2, \cdots, x_n\}$, tem-se: $f(x) = p_n(x) + p_{nodal}(x)q(x)$ em que o último termo $R(x) = p_{nodal}(x)q(x)$ corresponde ao erro da interpolação. Para calcular a forma de $q(x)$ a seguinte função é proposta:

$$Q(t) = f(t) - p_n(t) - p_{nodal}(t)q(x).$$

Sendo t um valor genérico do argumento da função que satisfaz a:

$$t \in \begin{cases} \{x, x_0, x_1, \cdots, x_n\} \text{ se } x < x_0 \\ \{x_0, x_1, \cdots, x_n\} \text{ se } x_0 \le x \le x_n \\ \{x_0, x_1, \cdots, x_n, x\} \text{ se } x > x_n \end{cases}.$$

[6] Charles Hermite (1822-1901).

Como $Q(t)=0$ para $(n+2)$ valores $[t=x_0, t=x_1, t=x_2, \cdots, t=x_n$ e $t=x]$ sua derivada se anula em pelo menos $(n+1)$ valores no intervalo, sua derivada segunda se anula em pelo menos n valores no intervalo, sua derivada terceira se anula em pelo menos $(n-1)$ valores no intervalo, e assim sucessivamente. Assim, a função $Q^{(k)}(t)$ se anula em pelo menos $(n+2-k)$ (k é a ordem da derivada) pontos no intervalo, então se $(n+2-k)=1 \Rightarrow k=(n+1)$, isto, é a derivada de ordem $(n+1)$ se anula em pelo menos um ponto no intervalo. Seja este valor $t=\xi$, mas $\dfrac{d^{n+1}p_n(t)}{dt^{n+1}}=0$, pois $p_n(t)$ é um polinômio de grau n e $p_{nodal}(t)=\prod_{j=0}^{n}(t-x_j)=t^{n+1}+\cdots \Rightarrow \dfrac{d^{n+1}p_{nodal}(t)}{dt^{n+1}}=(n+1)!$, então:

$$\left.\frac{d^{n+1}Q(t)}{dt^{n+1}}\right|_{t=\xi}=0=\left.\frac{d^{n+1}f(t)}{dt^{n+1}}\right|_{t=\xi}-(n+1)q(x)\Rightarrow q(x)=\frac{1}{(n+1)!}\left.\frac{d^{n+1}f(t)}{dt^{n+1}}\right|_{t=\xi}.$$

Resultando em:

$$f(x)=p_n(x)+p_{nodal}(x)\frac{1}{(n+1)!}\left.\frac{d^{n+1}f(t)}{dt^{n+1}}\right|_{t=\xi}.$$

Do procedimento recursivo descrito na Seção 3.3, tem-se:

$$a_{n+1}=f[x,x_n,x_{n-1},\cdots,x_2,x_1,x_0]=\frac{f(x)-p_n(x)}{(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})(x-x_n)},$$

identificando:$(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})(x-x_n)=p_{nodal}(x)$, obtém-se:

$$f(x)=p_n(x)+p_{nodal}(x)f[x,x_n,x_{n-1},\cdots,x_2,x_1,x_0].$$

Comparando esta expressão com a expressão anterior, chega-se à conclusão que:

$$f[x,x_n,x_{n-1},\cdots,x_2,x_1,x_0]=\frac{1}{(n+1)!}\left.\frac{d^{n+1}f(t)}{dt^{n+1}}\right|_{t=\xi}.$$

Procedimento semelhante pode ser feito com o método de interpolação de Lagrange, após ser aplicado o procedimento descrito na Seção 3.4 relativo ao cálculo da matriz $\mathbf{G}$ que contém, em cada linha i, os valores dos coeficientes da expansão de $p_n(x)$ em torno do ponto nodal x_i. Isto permite inferir o valor da função $q(x)$ para um valor específico de x que é igual ao valor do coeficiente de ordem $(n+1)$ da interpolação polinomial resultante da adição deste valor como novo ponto nodal, assim:

$$q(x)=a_{n+1}=\frac{f(x)-p_n(x)}{p_{nodal}(x)} \text{ sendo } p_n(x)=\sum_{j=0}^{n}\mathbf{G}_{m,j}(x-x_m)^j$$

Deve-se ressaltar que este valor independe de $m\,\{0\leq m\leq n\}$ e que o valor deste coeficiente é igual ao valor obtido pelo método das diferenças divididas de Newton. Desta maneira, demonstra-se a equivalência do método de interpolação baseado nas diferenças divididas de Newton e do método de interpolação de Lagrange, pois, nos dois métodos, é possível avaliar o erro da interpolação sem o conhecimento da forma explícita da função a ser interpolada, sendo desnecessária a diferenciação sucessiva da mesma.

A expressão do erro na interpolação de Hermite é obtida de forma semelhante ao da determinação do erro na interpolação polinomial de Lagrange, resultando em:

$$f(x)=p_{2n+1}(x)+p_{nodal}^2(x)\frac{1}{(2n+2)!}\left.\frac{d^{2n+2}f(t)}{dt^{2n+2}}\right|_{t=\xi}.$$

3.6 Critério de Minimização do Erro Quadrático Médio

Na determinação dos coeficientes da aproximação em problemas em que a função $f(x)$ é conhecida, novos procedimentos podem ser desenvolvidos baseados na análise da integral do quadrado do erro, descrita pela equação:

$$J(\mathbf{c}) = \frac{1}{b-a}\int_a^b \left[f(x) - \sum_{i=0}^{n} c_i x^i\right]^2 dx.$$

O desenvolvimento dos métodos é facilitado se o intervalo $[a,b]$ for normalizado por um dos procedimentos descritos anteriormente.

- Mudança de variável para $t = \dfrac{2x-(a+b)}{b-a}$ resultando em:

$$J_1(\mathbf{c}) = \frac{1}{2}\int_{-1}^{+1} \left[f[x(t)] - \sum_{i=0}^{n} c_i t^i\right]^2 dt.$$

- Mudança de variável para $t = \dfrac{x-a}{b-a}$ resultando em:

$$J_2(\mathbf{c}) = \int_{0}^{+1} \left[f[x(t)] - \sum_{i=0}^{n} c_i t^i\right]^2 dt.$$

Os valores dos coeficientes são determinados pela minimização dos valores dessas integrais, dando origem aos sistemas lineares.

- $\dfrac{\partial J_1(\mathbf{c})}{\partial c_j} = -\int_{-1}^{+1} t^j \left[f[x(t)] - \sum_{i=0}^{n} c_i t^i\right] dt \Rightarrow \sum_{i=0}^{n} \dfrac{1-(-1)^{i+j+1}}{i+j+1} c_i = \int_{-1}^{+1} t^j f[x(t)]dt.$

  $\mathbf{A}\cdot\mathbf{c} = \mathbf{b}$ sendo $b_j = \int_{-1}^{+1} t^j f[x(t)]dt$ e, como:

  $1-(-1)^{i+j+1} = \begin{cases} 2 \text{ para } i+j = 2k \\ 0 \text{ para } i+j = 2k+1 \end{cases}$ para $k = 0, 1, 2, 3, \cdots$, resulta:

  n par: $\mathbf{A} = 2\begin{bmatrix} 1 & 0 & \frac{1}{3} & 0 & \cdots & \frac{1}{n+1} \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & 0 \\ \frac{1}{3} & 0 & \frac{1}{5} & 0 & \cdots & \frac{1}{n+3} \\ 0 & \frac{1}{5} & 0 & \frac{1}{7} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n+1} & 0 & \frac{1}{n+3} & 0 & \cdots & \frac{1}{2n+1} \end{bmatrix}$, n ímpar: $\mathbf{A} = 2\begin{bmatrix} 1 & 0 & \frac{1}{3} & 0 & \cdots & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \frac{1}{3} & 0 & \frac{1}{5} & 0 & \cdots & 0 \\ 0 & \frac{1}{5} & 0 & \frac{1}{7} & \cdots & \frac{1}{n+4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \frac{1}{n+2} & 0 & \frac{1}{n+4} & \cdots & \frac{1}{2n+1} \end{bmatrix}$

- $\dfrac{\partial J_2(\mathbf{c})}{\partial c_j} = -2\int_{0}^{+1} t^j \left[f[x(t)] - \sum_{i=0}^{n} c_i t^i\right] dt \Rightarrow \sum_{i=0}^{n} \dfrac{1}{i+j+1} c_i = \int_{0}^{+1} t^j f[x(t)]dt.$

  $\mathbf{A}\cdot\mathbf{c} = \mathbf{b}$ sendo $b_j = \int_{0}^{+1} t^j f[x(t)]dt$ e

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots & \frac{1}{n+3} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \cdots & \frac{1}{n+4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \frac{1}{n+3} & \frac{1}{n+4} & \cdots & \frac{1}{2n+1} \end{bmatrix}$$

Para ilustrar este procedimento o Exemplo 3.1 é refeito.

■ **Exemplo 3.5** Aproximação Polinomial de Terceiro Grau da Função $f(x) = \dfrac{20x}{1+20x^2}$ no intervalo fechado $[0,1]$, intervalo já normalizado.

Polinômio interpolador de terceiro grau que minimiza a integral do quadrado do erro no intervalo, neste caso os coeficientes são determinados pela resolução do sistema linear:

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \int_0^{+1} \dfrac{20x}{1+20x^2} dx \\ \int_0^{+1} \dfrac{20x^2}{1+20x^2} dx \\ \int_0^{+1} \dfrac{20x^3}{1+20x^2} dx \\ \int_0^{+1} \dfrac{20x^4}{1+20x^2} dx \end{bmatrix} = \begin{bmatrix} 1,522261 \\ 0{,}69795 \\ 0{,}423887 \\ 0,298436 \end{bmatrix} \Rightarrow \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0{,}55402 \\ 11,746197 \\ -25,765618 \\ 14,734727 \end{bmatrix}$$

O valor da função objetivo, $J(\mathbf{c}) = \displaystyle\int_0^{+1} \left[f[x(t)] - \sum_{i=0}^{3} c_i t^i \right]^2 dt = 0,027025$. Usando o valor de $\mathbf{c} = \begin{bmatrix} 0 \\ 13,506 \\ -26,568 \\ 14,015 \end{bmatrix}$ como determinado no Exemplo 3.1, obtém-se o valor $J(\mathbf{c}) = 0,069678$, cerca de 2,6 vezes maior que o novo valor!

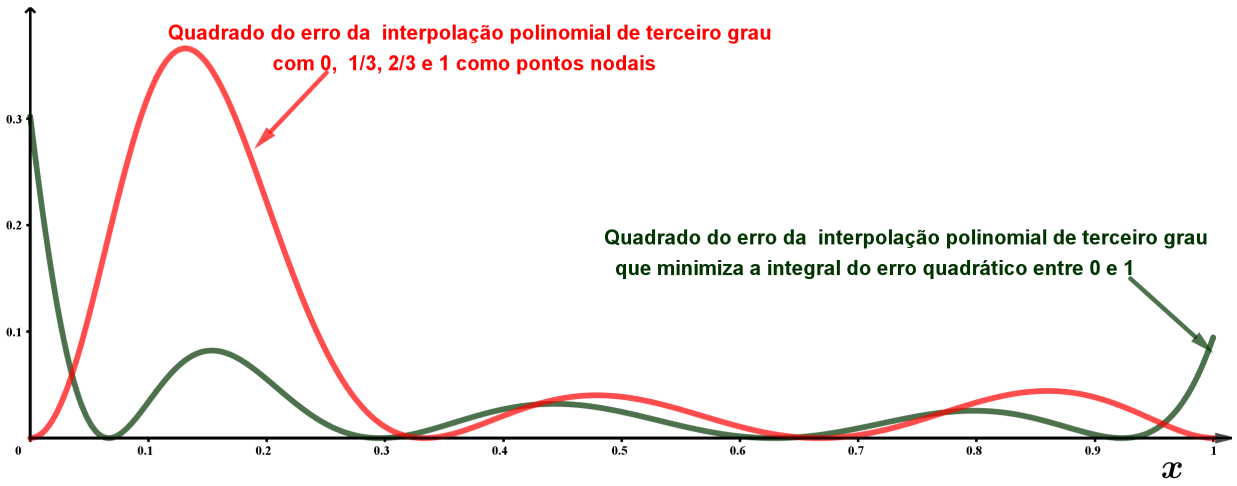


Figura 3.6: Quadrado dos erros de aproximações polinomiais de terceiro grau de $f(x) = \dfrac{20x}{1+20x^2}$.

■

Uma forma que facilita bastante este procedimento é a utilização de **Polinômios de Legendre**[7],

[7] Adrien-Marie Legendre (1752-1833).

$P_i(x)$, que apresenta a seguinte propriedade de *ortogonalidade*[8]:

$$\int_0^{+1} P_i(x)P_j(x)dx = \frac{1}{2i+1}\delta_{i,j} = \begin{cases} \dfrac{1}{2i+1} \text{ para } i=j \\ 0 \text{ para } i \neq j \end{cases} \quad \text{para } i,j = 0,\,1,\,2,\,3,\,\cdots.$$

Os quatro primeiros polinômios de Legendre são a seguir apresentados:
$P_0(x) = 1$
$P_1(x) = 2x-1$
$P_2(x) = 6x^2-6x+1$
$P_3(x) = 20x^3-30x^2+12x-1$

Estes polinômios são gerados a partir da equação de recorrência:

$$P_k(x) = \frac{(2k-1)P_{k-1}(x)-(k-1)P_{k-2}(x)}{k} \text{ para } k = 2,\,3,\,\cdots \text{ com } P_0(x)=1 \text{ e } P_1(x)=2x-1.$$

A aproximação polinomial de grau n é então expressa na forma:

$$p_n(x) = \sum_{i=0}^{n} a_i P_i(x) \text{ sendo } a_i = (2i-1)\int_0^{+1} P_i(x)f(x)dx.$$

Ou ainda, fazendo uso das raízes do Polinômio de Legendre de grau $(n+1)$ como pontos nodais da interpolação $\{x_0, x_1, \cdots, x_n\}$, temos como boa aproximação para $f(x)$, segundo o critério da minimização do erro médio quadrático:

$$p_n(x) = \sum_{j=0}^{n} \ell_j(x) f(x_j).$$

Deve-se enfatizar que este procedimento de minimização só pode ser aplicado se a função $f(x)$ for definida e conhecida em todo o intervalo.

3.7 Critério de Minimização do Erro Máximo

Na Seção 3.6 foi descrita uma forma de determinar os coeficientes da aproximação polinomial de uma função $f(x)$ conhecida pela minimização da integral do erro quadrático médio, descrita pela equação:

$$J(\mathbf{c}) = \frac{1}{b-a}\int_a^b \left[f(x) - \sum_{i=0}^{n} c_i x^i\right]^2 dx = \frac{1}{b-a}\int_a^b [R(x)]^2\,dx.$$

Sendo também indicado um procedimento que evita a resolução do sistema linear resultante através da expansão da aproximação polinomial em termos dos n primeiros polinômios de Legendre.

Outro critério de otimização que pode também ser empregado é o da minimização do módulo do erro máximo da aproximação no intervalo considerado. Isto pode ser feito minimizando os valores máximos do módulo do polinômio nodal, pois o módulo do erro da aproximação polinomial é descrito por:

$$|R(x)| = |p_{nodal}(x)|\frac{1}{(n+1)!}\left.\frac{d^{n+1}f(t)}{dt^{n+1}}\right|_{t=\xi}$$

Verificando-se que o módulo do erro diminui com o aumento do número de pontos nodais, aumenta com o aumento do módulo da derivada de ordem $(n+1)$ da função a ser aproximada e

[8] Na maioria dos textos, os polinômios de Legendre são definidos no intervalo $[-1,+1]$. Para adaptá-los ao intervalo $[0,+1]$ basta trocar seu argumento x por $(2x-1)$.

aumenta com o aumento do módulo do polinômio nodal. Assim, o valor máximo do erro, para uma aproximação polinomial de grau n somente poderia ser minimizado pela seleção adequada dos $(n+1)$ pontos nodais. O polinômio que apresenta os menores valores de máximos e mínimos é o **Polinômio de Chebyshev**[9] que nada mais é que o polinômio de grau n em $\cos(\theta)$ que expressa o cosseno múltiplo de um arco em termos do cosseno do arco, isto é $\cos(n\theta)$ para $0 \leq \theta \leq \pi$, ou seja, um componente da série de Cosseno de Fourier (ver Seção 2.5). No Capítulo 2 apresentaram-se as equações:

$$\cos[(k+1)\theta] = 2\cos(\theta)\cos(k\theta) - \cos[(k-1)\theta] \text{ para } k = 1, 2, \cdots$$

Definindo $T_k(\theta) = \cos(k\theta) \Rightarrow T_0(\theta) = 1$ e $T_1(\theta) = \cos(\theta)$, tem-se o procedimento recursivo:

$$T_{k+1}(\theta) = 2T_1(\theta)T_k(\theta) - T_{k-1}(\theta) \text{ para } k = 1, 2, \cdots \text{ com } T_0(\theta) = 1 \text{ e } T_1(\theta) = \cos(\theta).$$

Definindo a variável $x = \cos(\theta)$ e para $0 \leq \theta \leq \pi \Rightarrow -1 \leq x \leq +1$ tem-se:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \text{ para } k = 1, 2, \cdots \text{ com } T_0(x) = 1 \text{ e } T_1(x) = x.$$

O polinômio de Chebyshev $T_n(x)$ apresenta as seguintes propriedades:

1. $|T_n(x)| \leq 1$ no intervalo $-1 \leq x \leq +1$;
2. Suas raízes são $x_k = \cos\left[\dfrac{(2k-1)\pi}{2n}\right]$ para $k = 1, 2, \cdots n$;
3. Como $T_n(x) = \cos(n\theta)$ com $0 \leq \theta \leq \pi$, em vista de:

$$\int_{-1}^{+1} \cos(i\pi x)\cos(j\pi x)dx = 2\int_{0}^{+1} \cos(i\pi x)\cos(j\pi x)dx = \begin{cases} 0 \text{ para } i \neq j \\ 1 \text{ para } i = j \neq 0 \\ 2 \text{ para } i = j = 0 \end{cases},$$

Considerando $\theta = \pi x \Rightarrow \displaystyle\int_0^{\pi} \cos(i\theta)\cos(j\theta)d\theta = \begin{cases} 0 \text{ para } i \neq j \\ \dfrac{\pi}{2} \text{ para } i = j \neq 0 \\ \pi \text{ para } i = j = 0 \end{cases}$

e como $x = \cos(\theta) \Rightarrow dx = -\operatorname{sen}(\theta)d\theta = -\sqrt{1-x^2}d\theta$ ou seja $d\theta = -\dfrac{dx}{\sqrt{1-x^2}}$, resultando finalmente em:

$$\int_{-1}^{+1} \frac{1}{\sqrt{1-x^2}} T_i(x)T_j(x)dx = \begin{cases} 0 \text{ para } i \neq j \\ \dfrac{\pi}{2} \text{ para } i = j \neq 0 \\ \pi \text{ para } i = j = 0 \end{cases}.$$

O que caracteriza os polinômios de Chebyshev como uma *família* de polinômios ortogonais no intervalo $[-1, +1]$ em relação à função peso $w(x) = \dfrac{1}{\sqrt{1-x^2}}$.

[9] Pafnuty Lvovich Chebyshev (1821-1894).

Os dez primeiros polinômios de Chebyshev são listados a seguir:

$$\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1 \\
T_7(x) &= 64x^7 - 112x^5 + 56x^3 - 7x \\
T_8(x) &= 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1 \\
T_9(x) &= 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x
\end{aligned}$$

Outra propriedade importante dos polinômios de Chebyshev é a possibilidade de todas as potências da variável x poderem ser expressas como uma combinação linear dos mesmos, conforme listado a seguir para as dez primeiras potências:

$$\begin{aligned}
1 &= T_0(x) \\
x &= T_1(x) \\
x^2 &= \frac{T_2(x) + T_0(x)}{2} \\
x^3 &= \frac{T_3(x) + 3T_1(x)}{4} \\
x^4 &= \frac{T_4(x) + 4T_2(x) + 3T_0(x)}{8} \\
x^5 &= \frac{T_5(x) + 5T_3(x) + 10T_1(x)}{16} \\
x^6 &= \frac{T_6(x) + 6T_4(x) + 15T_2(x) + 10T_0(x)}{32} \\
x^7 &= \frac{T_7(x) + 7T_5(x) + 21T_3(x) + 35T_1(x)}{64} \\
x^8 &= \frac{T_8(x) + 8T_6(x) + 28T_4(x) + 56T_2(x) + 35T_0(x)}{128} \\
x^9 &= \frac{T_9(x) + 9T_7(x) + 36T_5(x) + 84T_3(x) + 126T_1(x)}{256}
\end{aligned}$$

As formas dos 6 primeiros polinômios de Chebyshev são mostradas na Figura 3.7.

Considerando que o coeficiente do termo x^n em $T_n(x)$ é igual a 2^{n-1} sua forma *normalizada*, com o coeficiente do termo x^n igual a 1, designado por $t_n(x) = \frac{T_n(x)}{2^{n-1}}$ apresenta a propriedade $|t_n(x)| \leq \frac{1}{2^{n-1}}$.

Na Figura 3.8 são confrontados os polinômios nodais usando pontos nodais igualmente espaçados com polinômios nodais iguais aos polinômios de Chebyshev normalizados, para $n = 2, 3, 4$ e 5.

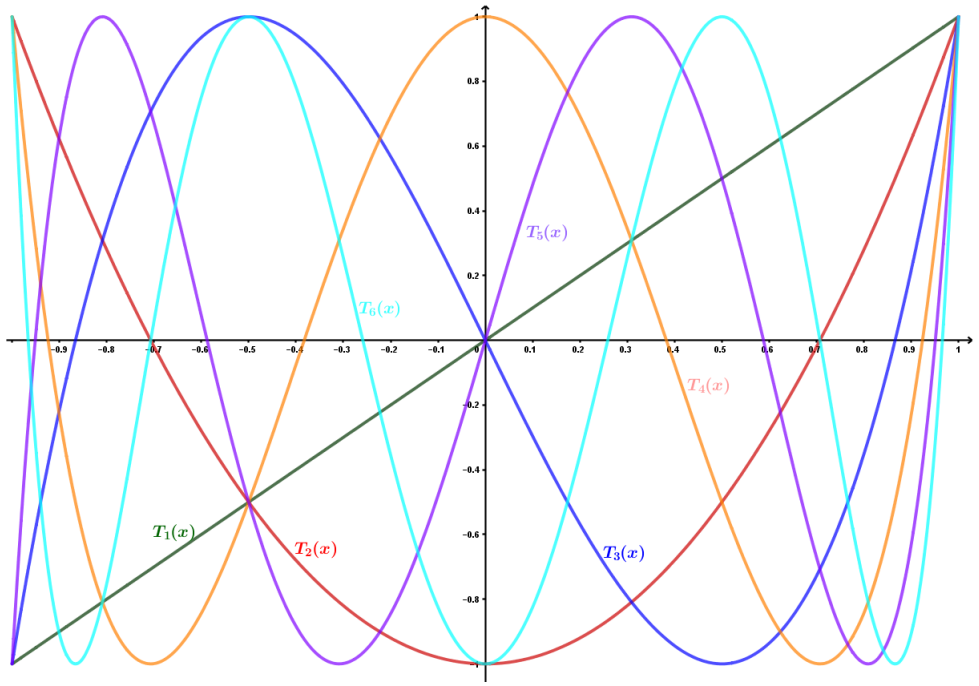


Figura 3.7: Polinômios de Chebyshev.

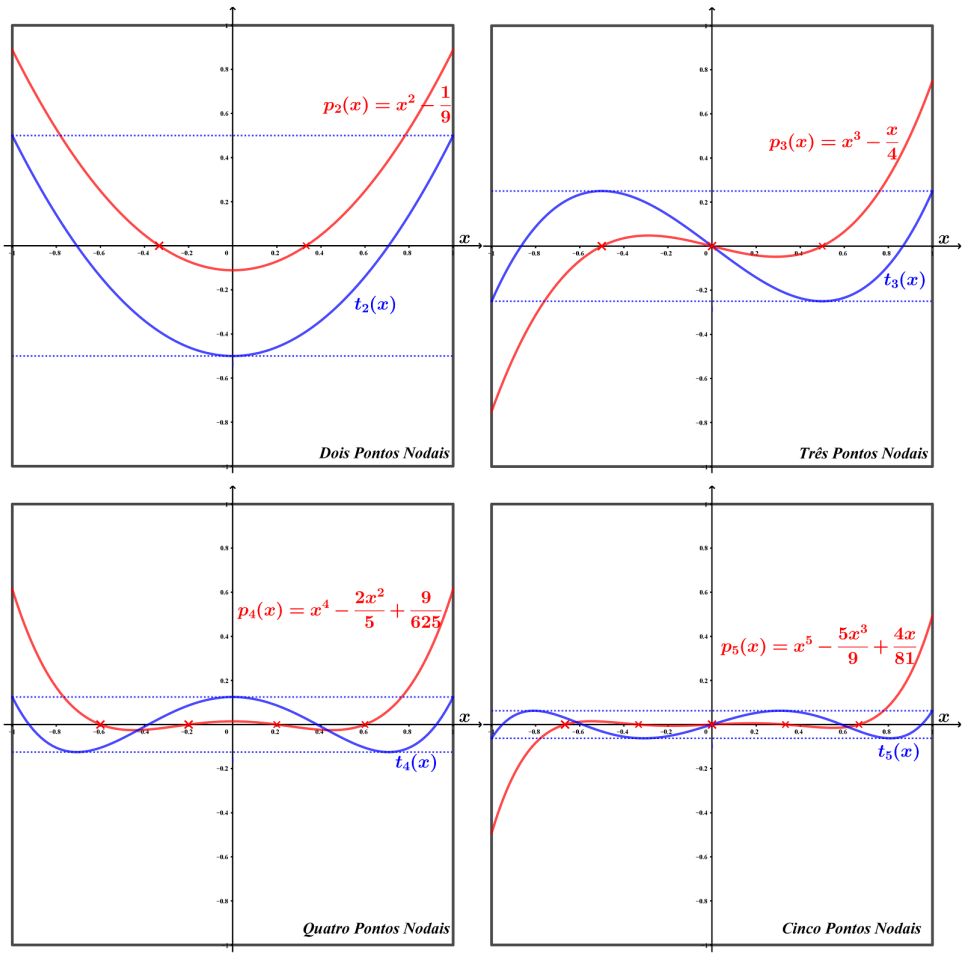


Figura 3.8: Interpolação de Chebyshev *versus* interpolação com pontos igualmente espaçados.

A análise dos gráficos da Figura 3.8 permite verificar que adotar o polinômio de Chebyshev normalizado como polinômio nodal, o que equivale a escolher como pontos nodais as raízes do polinômio de Chebyshev de grau $(n+1)$, faz com que os máximos valores dos módulos dos erros sejam os menores possíveis. Este é o chamado princípio do *mini-max*, traduzido pelo mínimo dos máximos dos módulos dos erros da aproximação polinomial de grau n.

Este fato é reforçado pela análise da interpolação polinomial da **função de Runge**[10], ilustrada na Figura 3.9:

$$f(x) = \frac{1}{1+25x^2}$$

O comportamento oscilatório das interpolações polinomiais com pontos nodais igualmente espaçados, é conhecido em análise numérica como o **Fenômeno de Runge**. Esse fenômeno foi caracterizado por Runge ao analisar o comportamento dos erros em aproximações polinomiais de diversas funções.

O valor do módulo do erro da interpolação pode também ser avaliado através da integral do quadrado do polinômio nodal, assim os pontos nodais, representados pelo vetor $\mathbf{x}$, também podem ser selecionados pela minimização da função:

$$J(\mathbf{x}) = \int_{-1}^{+1} \left[\prod_{k=0}^{n}(x-x_k)\right]^2 dx = \int_{-1}^{+1} p_{nodal}^2(x,\mathbf{x})dx$$

Em vista da propriedade de ortogonalidade dos polinômios de Legendre (definido no intervalo de ortogonalidade usual $[-1,+1]$), tem-se:

$$\int_{-1}^{+1} P_i(x)P_j(x)dx = \frac{2}{2i+1}\delta_{i,j}$$

e, em decorrência da propriedade de ortogonalidade, tem-se:

$$\int_{-1}^{+1} x^k p_i(x)dx = \begin{cases} 0 \text{ para } k < i \\ \dfrac{2}{(2i+1)C_i^2} \end{cases}$$

Sendo C_i o coeficiente de x^i em $P_i(x)$ e $p_i(x) = \dfrac{P_i(x)}{C_i}$, o polinômio de Legendre de grau i *normalizado*. Os coeficientes C_i são obtidos pela recorrência $C_i = \dfrac{2i-1}{i}C_{i-1}$ com $C_0 = 1$ para $i = 1, 2, 3, \cdots$, os oito primeiros valores destes coeficientes são listados a seguir.

k	1	2	3	4	5	6	7	8
C_k	1	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{35}{8}$	$\frac{63}{8}$	$\frac{231}{16}$	$\frac{429}{16}$	$\frac{6435}{128}$

Deste modo, adotando como polinômio nodal o polinômio de Legendre normalizado de grau $(n+1)$, tem-se:

$$\frac{\partial J(\mathbf{x})}{\partial x_k} = \int_{-1}^{+1} \left[\prod_{j=0\neq k}^{n}(x-x_j)\right] p_{nodal}(x,\mathbf{x})dx$$

Como $\prod_{j=0\neq k}^{n}(x-x_j)$ é um polinômio em x de grau n e $p_{nodal}(x,\mathbf{x}) = p_{n+1}(x)$ [polinômio de Legendre normalizado de grau $(n+1)$], essa integral é nula para $k = 0, 1 \cdots, n$ assegurando que as $(n+1)$ raízes do polinômio de Legendre são os pontos nodais que minimizam a função $J(\mathbf{x})$.

[10] Carl David Tolmé Runge (1856-1927).

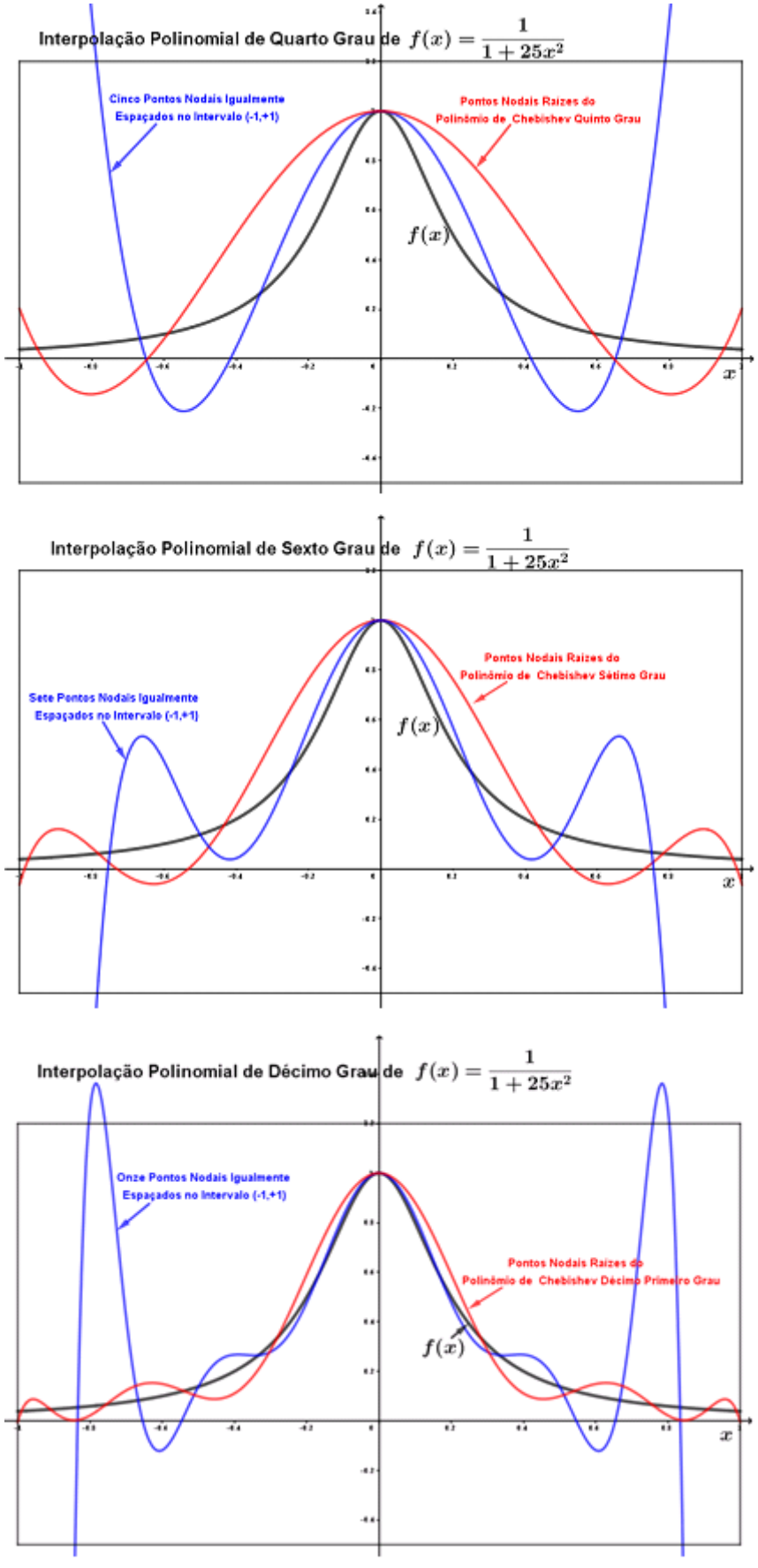


Figura 3.9: Interpolação de Chebyshev *versus* interpolação com pontos igualmente espaçados da função de Runge.

Para efeito de comparação, o valor de $J(\mathbf{x})$ é calculado para três casos: (a) os pontos nodais são as raízes do polinômio de Legendre de grau $(n+1)$; (b) os pontos nodais são as raízes do polinômio de Chebyshev de grau $(n+1)$ e (c) os pontos nodais são $(n+1)$ pontos igualmente espaçados no interior do intervalo $[-1,+1]$.

Número de pontos nodais	2	3	4	5
Raízes de $P_{Legendre}$	0,1778	0,0457	0,0116	0,0029
Raízes de $P_{Chebyshev}$	0,2333	0,0607	0,0154	0,0039
Igualmente espaçados	0,2765	0,1274	0,0619	0,0310

3.8 Telescopagem de Séries

Outra aplicação importante dos polinômios de Chebyshev é a **Telescopagem de Séries de Potências**, que consiste em expressar as sucessivas potências da variável x em termos dos polinômios de Chebyshev. Se a variável x estiver contida em um intervalo distinto do intervalo de ortogonalidade do polinômio de Chebyshev, $[-1,+1]$, a mesma deve ser normalizada para estar contida nesse intervalo, para isto aplica-se a mudança de variável $x \to z$ (como já descrito no início deste capítulo): $x = \dfrac{a+b}{2} + \dfrac{a-b}{2}\, z$. Por simplicidade, os desenvolvimentos seguintes são efetuados considerando que o argumento da função a ser aproximada já se encontra normalizado.

A explanação do procedimento de telescopagem será feita através de alguns exemplos clássicos.

■ **Exemplo 3.6** Telescopagem da função exponencial $f(x) = e^x$ com $x_0 = 0$.

Busca-se neste exemplo a aproximação polinomial de menor grau possível que apresente o módulo inferior a 2×10^{-3}. Neste caso a aproximação por polinômio de Taylor de quinto grau é:

$$p_5(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}$$

Que apresenta $R_5(x) = \dfrac{e^\xi}{720} \Rightarrow |R_5(x)| \leq \dfrac{e^\xi}{720} = 3,8 \times 10^{-3}$, podendo ser maior do que 2×10^{-3}. Busca-se assim o polinômio de Taylor de sexto grau.

$$p_6(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720}$$

Que apresenta $R_6(x) = \dfrac{e^\xi}{5040} \Rightarrow |R_6(x)| \leq \dfrac{e}{5040} = 5,4 \times 10^{-4} < 2 \times 10^{-3}$ satisfazendo à restrição imposta.

Verifica-se, a seguir, as seguintes possibilidades:

(a) Aproximação dos termos x^4, x^5 e x^6 por potências inferiores:

$$x^4 = \frac{T_4(x) + 4T_2(x) + 3T_0(x)}{8} \approx \frac{4T_2(x) + 3T_0(x)}{8} \Rightarrow \frac{x^4}{24} \approx \frac{4T_2(x) + 3T_0(x)}{24 \cdot 8}$$

Como $|Erro| \leq \dfrac{1}{8 \cdot 24} \approx 5,2 \times 10^{-3} > 2 \times 10^{-3}$, o que elimina tal possibilidade.

(b) Aproximação dos termos x^5 e x^6 por potências inferiores:

$$x^5 = \frac{T_5(x) + 5T_3(x) + 10T_1(x)}{16} \approx \frac{5T_3(x) + 10T_1(x)}{16} \Rightarrow \frac{x^5}{120} \approx \frac{5T_3(x) + 10T_1(x)}{16 \cdot 120}$$

Como $|Erro| \leq \dfrac{1}{16 \cdot 120} \approx 5,2 \times 10^{-4} < 2 \times 10^{-3}$, não elimina tal possibilidade.

$$x^6 = \frac{T_6(x) + 6T_4(x) + 15T_2(x) + 10T_0(x)}{32} \approx \frac{6T_4(x) + 15T_2(x) + 10T_0(x)}{32}$$

$$\frac{x^6}{720} \approx \frac{6T_4(x)+15T_2(x)+10T_0(x)}{32\cdot 720}$$

Como $|Erro| \leq \dfrac{1}{32\cdot 720} \approx 4,3\times 10^{-5} < 2\times 10^{-3}$, não elimina tal possibilidade.

Somando os três módulos de erro $5,4\times 10^{-4}+5,2\times 10^{-4}+4,3\times 10^{-5} \approx 1,1\times 10^{-3} < 2\times 10^{-3}$ satisfazendo portanto à restrição.

A forma da aproximação é então composta:

$$\frac{x^5}{120}+\frac{x^6}{720} \approx \frac{5T_3(x)+10T_1(x)}{16\cdot 120}+\frac{6T_4(x)+15T_2(x)+10T_0(x)}{32\cdot 720} = \frac{x^4}{480}+\frac{x^3}{96}-\frac{x^2}{1280}-\frac{x}{384}+\frac{1}{23040}$$

$$p_6(x) = 1+x+\frac{x^2}{2}+\frac{x^3}{6}+\frac{x^4}{24}+\frac{x^5}{120}+\frac{x^6}{720} \approx \frac{23041}{23040}+\frac{383x}{384}+\frac{639x^2}{1280}+\frac{17x^3}{96}+\frac{7x^4}{160}$$

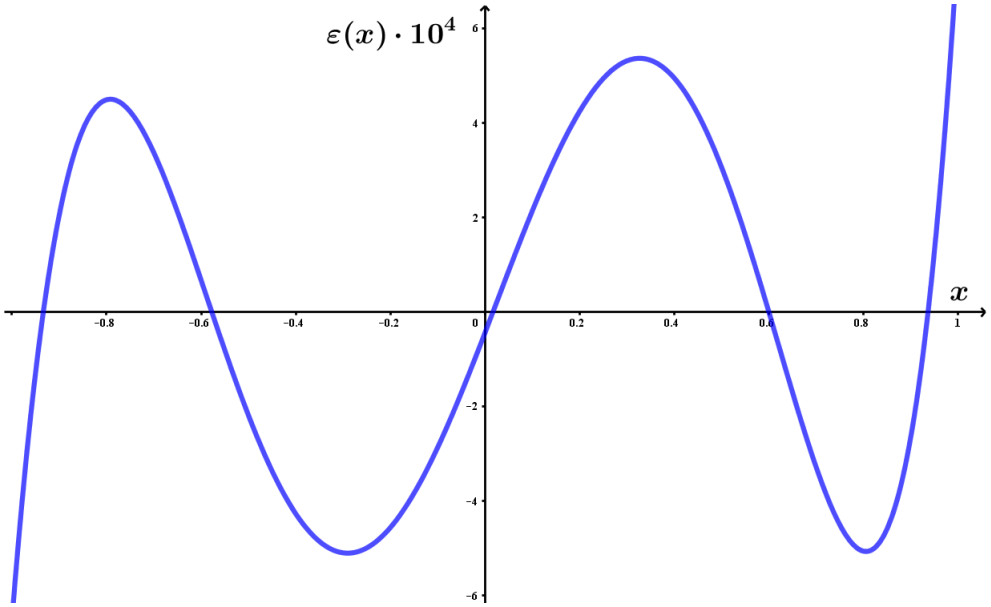


Figura 3.10: Erro da aproximação de e^x por polinômio de sexto grau telescopado em polinômio de quarto grau.

■

Uma forma mais rápida de obter esta aproximação (desde que se tenha um bom *pacote* de integração numérica) baseia-se na ortogonalidade dos polinômios de Chebyshev no intervalo considerado, assim uma função $f(x)$ contínua e definida no intervalo pode ser expandida na forma:

$$f(x) \approx p_n(x) = \sum_{k=0}^{n} c_k T_k(x)$$

Sendo: $c_0 = \dfrac{1}{\pi}\displaystyle\int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}}dx = \frac{1}{\pi}\int_0^{\pi} f[\cos(\theta)]d\theta$

$c_k = \dfrac{2}{\pi}\displaystyle\int_{-1}^{+1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}dx = \frac{1}{\pi}\int_0^{\pi} f[\cos(\theta)]\cos(k\theta)d\theta$ para $k = 1, 2, 3, \cdots, n.$

O cômputo dos coeficientes por integração na variável θ é preferível devido à singularidade da função $w(x) = \dfrac{1}{\sqrt{1-x^2}}$ nos dois limites da integral.

Ou ainda, fazendo uso das raízes do Polinômio de Chebyshev de grau $(n+1)$ como pontos nodais da interpolação $\{x_0, x_1, \cdots, x_n\}$, temos como boa aproximação para $f(x)$, segundo o critério da minimização do erro máximo:

$$p_n(x) = \sum_{j=0}^{n} \ell_j(x) f(x_j).$$

■ **Exemplo 3.7** Telescopagem da função cosseno $f(x) = \cos(x)$ com $x_0 = 0$

Da mesma forma que no exemplo anterior, busca-se neste exemplo a aproximação polinomial de menor grau possível que apresente o módulo inferior a 2×10^{-3}. Neste caso a aproximação por polinômio de Taylor de sexto grau é:

$$p_6(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720}$$

Que apresenta (pelo fato da correspondente série de Taylor ser uma série cujos termos têm sinais alterados o que implica no módulo do erro de truncamento ser menor ou igual ao máximo módulo do primeiro termo não considerado) $|R_6(x)| \leq \dfrac{1}{8!} = 2{,}5 \times 10^{-5} < 2 \times 10^{-3}$ inferior à acurácia desejada.

$x^6 = \dfrac{T_6(x) + 6T_4(x) + 15T_2(x) + 10T_0(x)}{32} \approx \dfrac{6T_4(x) + 15T_2(x) + 10T_0(x)}{32}$

Então: $\dfrac{x^6}{720} \approx \dfrac{6T_4(x) + 15T_2(x) + 10T_0(x)}{720 \cdot 32} = \dfrac{1}{23040} - \dfrac{x^2}{1280} + \dfrac{x^4}{480}$

Com $|Erro| \leq \dfrac{1}{720 \cdot 32} = 4{,}34 \times 10^{-5}$ que somado ao erro de truncamento

$|Erro|_{total} \leq \dfrac{1}{8!} + \dfrac{1}{720 \cdot 32} = 6{,}82 \times 10^{-5}$

A forma da aproximação é então composta:

$$\cos(x) \approx p_4(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \left[\frac{1}{23040} - \frac{x^2}{1280} + \frac{x^4}{480}\right] = \frac{23039}{23040} - \frac{639x^2}{1280} + \frac{19x^4}{480}$$

com $|Erro| \leq 6{,}82 \times 10^{-5}$.

Por expansão de $f(x) = \cos(x)$ em série de polinômios de Chebyshev, obtém-se:

$$\cos(x) \approx c_0 T_0(x) + c_1 T_2(x) + c_2 T_4(x)$$

Sendo:

$$\begin{aligned} c_0 &= \frac{1}{\pi}\int_0^{\pi} \cos(\cos(\theta))d\theta = 0{,}7651977 \\ c_1 &= \frac{2}{\pi}\int_0^{\pi} \cos(\cos(\theta))\cos(2\theta)d\theta = -0{,}229807 \\ c_2 &= \frac{2}{\pi}\int_0^{\pi} \cos(\cos(\theta))\cos(4\theta)d\theta = 0{,}00495328 \end{aligned}$$

A forma da aproximação é então composta:

$\cos(x) \approx p_4(x) = 0{,}7651977 - 0{,}229807(2x^2 - 1) + 0{,}00495328(8x^4 - 8x^2 + 1) =$

$= 0,9999580 - 0,4992402x^2 + 0,0396262x^4$.

Com $|Erro| \leq 4,19 \times 10^{-5} = \left| c_3 = \frac{2}{\pi}\int_0^{\pi} \cos(\cos(\theta))\cos(6\theta)d\theta \right|$, módulo do primeiro termo não considerado na expansão em série de polinômios de Chebyshev.

Obtendo a aproximação polinomial de grau 4 usando como pontos nodais as raízes de $T_5(x)$, obtém-se o seguinte resultado:

$$p_4(x) = \sum_{j=0}^{4} \ell_j(x) f(x_j) = 1 - 0,4995756x^2 + 0,0399612x^4.$$

■

Para calcular a aproximação de grau n de uma função $f(x)$ contínua e definida no intervalo $[-1,+1]$ expandida na forma:

$$f(x) \approx p_n(x) = \sum_{k=0}^{n} c_k T_k(x),$$

um procedimento recursivo, de simples implementação e semelhante ao método de Horner, foi proposto por Clenshaw (1955), traduzido pelo algoritmo:

$a_{n+2} \longleftarrow 0$

$a_{n+1} \longleftarrow 0$

Para $i = n, n-1, \cdots, 0,$ faça

$\quad a_i \longleftarrow c_i - a_{i+2} + 2x\, a_{i+1}$

$p_n \longleftarrow a_0 - a_1\, x.$

A telescopagem também pode ser realizada através de sucessivas reduções de grau do polinômio até a acurácia desejada usando o polinômio Chebyshev normalizado (ou *mônico*):

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

em que c_n é o coeficiente da maior potência de $p_n(x)$ e

$$|p_n(x) - p_{n-1}(x)| = |c_n t_n(x)| \leq \frac{c_n}{2^{n-1}}.$$

■ **Exemplo 3.8** Retomando o Exemplo 3.6 a partir da aproximação $p_6(x)$ da função exponencial, e aplicando o procedimento sequencial resulta em:

$$p_5(x) = p_6(x) - c_6\, t_6(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} - \frac{1}{720}\left(\frac{32x^6 - 48x^4 + 18x^2 - 1}{32}\right)$$

$$p_5(x) = \frac{23041}{23040} + x + \frac{639x^2}{1280} + \frac{x^3}{6} + \frac{21x^4}{480} + \frac{x^5}{120}$$

$$p_4(x) = p_5(x) - c_5\, t_5(x) = p_5(x) - \frac{1}{120}\left(\frac{16x^5 - 20x^3 + 5x}{16}\right) = \frac{23041}{23040} + \frac{383x}{384} + \frac{639x^2}{1280} + \frac{17x^3}{96} + \frac{7x^4}{160}.$$

■

3.9 Problemas Propostos

Problema 3.1 Busque uma expressão de segundo grau e outra de terceiro grau que *melhor* aproximam a função x^4 no intervalo $2 \leq x \leq 8$. Analise e discuta seus resultados confrontado-os graficamente.

Problema 3.2 Aproxime a função e^x no intervalo $0 \leq x \leq 2$ por um polinômio em x de menor grau possível que apresente o módulo do erro inferior a 10^{-4}. Confirme seu resultado representando a curva do erro da aproximação no intervalo considerado.

Problema 3.3 Hougen e Watson (1955) sugerem a seguinte expressão empírica para o cálculo do calor específico molar do nitrogênio gasoso $C_p(T) = 6,3 + 1,82 \times 10^{-3}T - 0,345 \times 10^{-6}T^2$ em que $C_p : cal/gmol/K$ e T : Kelvin. Na faixa de 300 a 2100 K, o erro máximo do calor específico calculado por esta expressão é de 1,2 %.

(a) determine a aproximação linear de C_p que minimiza o máximo do erro adicional na faixa de 1000 a 2000 K;

(b) Calcule o erro percentual máximo da aproximação proposta em (a) na mesma faixa de temperatura.

Problema 3.4 Para o cálculo da viscosidade do orto-xileno propõe-se o emprego da seguinte expressão: $\ln[\mu(T)] = -3,332 + \dfrac{1,039 \times 10^3}{T} - 1,768 \times 10^{-3}T + 1,076 \times 10^{-6}T^2$, em que T é a temperatura em Kelvin e μ é a viscosidade em centipoise, tal expressão é válida na faixa $245 \leq T \leq 620$ K. Obtenha a aproximação linear de que apresente o menor valor do módulo máximo do erro na faixa de 300 a 500 K. Calcule o valor máximo do módulo do erro da aproximação linear obtida nesta mesma faixa de temperatura.

Problema 3.5 A variação do coeficiente de expansão térmica do alumínio na faixa de 0 a 100 °C é dada por: $k(T) = 0,22 \times 10^{-4}T + 0,009 \times 10^{-6}T^2$ com T em °C.

(a) aproxime $k(T)$ por uma constante na mesma faixa de 0 a 100 °C, de modo que o módulo do erro máximo seja o menor possível;

(b) Calcule o valor médio $\bar{k} = \dfrac{1}{100}\displaystyle\int_0^{100} k(T)dT$, o valor médio aritmético, $k_{medio} = \dfrac{k(0)+k(100)}{2}$ e, através da comparação entre o valor obtido no item (a) e os dois valores médios, sugira qual valor é o mais apropriado.

Problema 3.6 Nas tabelas seguintes apresentam-se os valores das condutividades térmica do CO_2 e da viscosidade do etileno glicol líquido a várias temperaturas:

T (°F)	32	212	392	572
k (BTU/hr/ft/°F)	0,0085	0,0133	0,0181	0,0228

T (°F)	0	50	100	150	200
μ (lb/ft/hr)	242,00	82,10	30,50	12,60	5,57

Determine, em cada caso, o polinômio interpolador de menor grau possível que assegura um erro relativo (em módulo) inferior a 1% na faixa tabelada. No caso da viscosidade do etileno glicol aplique o polinômio interpolador ao $\ln(\mu)$.

Problema 3.7 No Problema 3.6, comentou-se que a viscosidade é bem representada por uma função exponencial da temperatura, isto é $\ln(\mu)$ é bem interpolado por uma função polinomial de T. Utilizando os valores da viscosidade do etileno glicol líquido em várias temperaturas (tabelados no Problema 3.6, construiu-se a Tabela de Diferenças Divididas abaixo:

k	T	$\ln(\mu)$	Δ_1	Δ_2	Δ_3	Δ_4
0	0	5,489				
			-0,022			
1	50	4,408		$1,816 \cdot 10^{-5}$		
			-0,020		$2,052 \cdot 10^{-6}$	
2	100	3,418		$2,124 \cdot 10^{-5}$		$-3,590 \cdot 10^{-10}$
			-0,018		$-5,127 \cdot 10^{-8}$	
3	150	2,534		$1,355 \cdot 10^{-5}$		
			-0,016			
4	200	1,717				

Os baixos valores de Δ_3 e Δ_4 indicam que a função $\ln(\mu)$ pode ser bem aproximada por uma função parabólica de T. Utilizando a Tabela de Diferenças acima calcule o polinômio interpolador de segundo grau [$\ln(\mu)$ *versus* T] e verifique os erros relativos do cálculo de μ nos pontos não utilizados na interpolação.

Problema 3.8 A tabela seguinte mostra a dependência da pressão parcial da amônia com a temperatura a diferentes concentrações.

Temperatura	Concentração percentual molal da amônia					
(°F)	0	10	20	25	30	35
60	0,26	1,42	3,51	5,55	8,65	13,22
80	0,51	2,43	5,85	9,06	13,86	20,61
100	0,95	4,05	9,34	14,22	21,32	31,16
140	2,89	9,98	21,49	31,54	45,73	64,78
180	7,51	21,65	44,02	62,68	88,17	121,68
220	17,19	42,47	81,91	113,81	156,41	211,24
250	29,83	66,67	124,08	169,48	229,62	305,60

Por interpolação linear *dupla* (envolvendo as duas variáveis independentes - temperatura e concentração), calcule as pressões parciais nos casos listados abaixo.

T (°C)	126,5	126,5	126,5	60,0	237,5	237,5
Concentração Molal (%)	28,8	6,7	25,0	15,0	17,6	35,0

Problema 3.9 O logaritmo neperiano de x pode ser determinado através da seguinte expansão em série de potências: $\ln(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i}$. Obtenha a aproximação parabólica de $\ln(1+x)$ que apresente o menor valor do módulo do erro no intervalo $0 \leq x \leq 1$. Analise e discuta os resultados confrontado-os graficamente no intervalo pertinente.

Problema 3.10 Busque uma expressão de quarto grau que melhor aproxime a função $\frac{1}{x^2}$ no intervalo $2 \leq x \leq 6$. Analise e discuta os resultados confrontado-os graficamente no intervalo pertinente.

Problema 3.11 Considere a função de Runge $f(x) = \frac{1}{1+25x^2}$ definida no intervalo $[-1,+1]$. Aproxime esta função por um polinômio de quarto grau em x que apresente os menores valores

dos máximos dos desvios. Baseado no fato da função $f(x)$ ser uma função par em x aproxime-a por uma função parabólica em $u = x^2$ que apresente no intervalo $0 \leq u \leq 1$ os menores valores dos máximos dos módulos dos desvios. Discuta e compare as duas aproximações obtidas.

Problema 3.12 Na tabela abaixo apresentam-se valores da viscosidade dinâmica da água a várias temperaturas.

T (°C)	0	5	10	20	30	40	50
$\mu \left[\frac{N \cdot s}{m^3}\right] \cdot 10^3$	1,787	1,519	1,307	1,002	0,798	0,653	0,547

Determine um polinômio interpolador de segundo grau que assegure um erro relativo inferior a 4,00 % em toda a faixa tabelada de T.

Problema 3.13 A função $\sqrt[3]{1+x}$, no intervalo $-1 \leq x \leq +1$, pode ser determinada através da seguinte expansão em série de potências: $\sqrt[3]{1+x} = 1 + \frac{x}{3} - \frac{2}{3 \cdot 6}x^2 + \frac{2 \cdot 5}{3 \cdot 6 \cdot 9}x^3 - \frac{2 \cdot 5 \cdot 8}{3 \cdot 6 \cdot 9 \cdot 12}x^4 + \cdots$.

Proponha uma aproximação parabólica de que apresente o menor valor do módulo do erro em todo domínio $-1 \leq x \leq +1$. Calcule o valor máximo do módulo do erro da aproximação, indicando o ponto em que ocorre.

4. Resolução Numérica de Equações em uma Variável

4.1 Introdução

Um modelo matemático bastante estudado na Engenharia Química é o modelo do reator químico contínuo de mistura perfeita (*Continuous Stirred Tank Reactor* ***CSTR***) não-isotérmico, sua não linearidade tem sido objeto de grande número de trabalhos científicos conforme reportado no livro de Aris (1999)[1]. Um diagrama simplificado deste tipo de reator é mostrado na Figura 4.1.

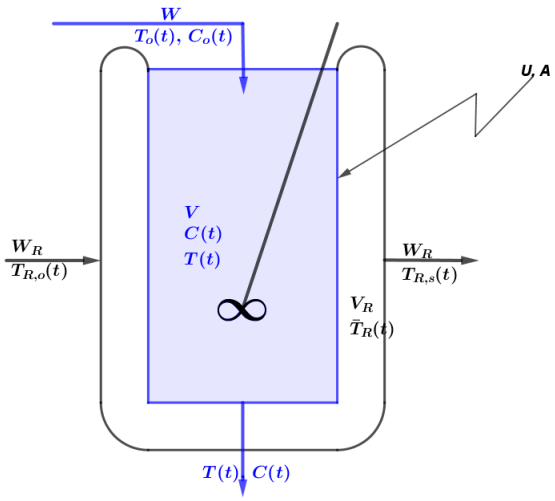


Figura 4.1: Diagrama de um reator químico contínuo de mistura perfeita.

Considerando as propriedades físicas constantes (ρ, C_p) e a ocorrência de uma reação irreversível, exotérmica e de primeira ordem, cuja expressão da velocidade da reação é descrita por: $r_A = k(T)C_A$ em que $k(T) = k_0\, e^{-\frac{E}{RT}}$, os balanços transientes de massa e de energia do modelo são descritos

[1] Rutherford Aris (1929-2005).

por:

$$\begin{aligned}
\frac{dV(t)}{dt} &= F_e(t) - F_s(t) \\
\frac{d[V(t)C_A(t)]}{dt} &= F_e(t)C_{Ae}(t) - F_s(t)C_A(t) - V(t)k[T(t)]C_A(t) \\
\rho V(t)C_p\frac{dT(t)}{dt} &= F_e(t)\rho C_p[T_e(t) - T(t)] + (-\Delta H_r)V(t)k[T(t)]C_A(t) - UA_t[T(t) - T_w(t)]
\end{aligned}$$

em que: V é o volume do meio reacional, F_e e F_s são as vazões volumétricas de entrada e de saída do reator, C_A é a concentração molar do reagente, T e T_w são as temperaturas do meio reacional e do fluido de refrigeração, ΔH_r é a entalpia da reação, U é o coeficiente global de transferência de calor e A_t é a área de troca térmica.

Os balanços estacionários de massa e de energia do modelo são então descritos por:

$$\begin{aligned}
F_e &= F_s \equiv F \\
\frac{F}{V}(C_{Ae} - C_A) &= k(T)C_A \\
\frac{F}{V}(T - T_e) + \frac{UA}{\rho C_p V}(T - T_w) &= \frac{(-\Delta H_r)k(T)C_A}{\rho C_p}
\end{aligned}$$

Definindo $\tau = \dfrac{V}{F}$ como o tempo de residência médio no reator, tem-se da equação de balanço de massa $C_A(T) = \dfrac{C_{Ae}}{1+k(T)\tau}$ que substituída no balanço de energia dá origem a:

$$(T - T_e) + \frac{UA}{F\rho C_p}(T - T_w) = \frac{(-\Delta H_r)C_{Ae}k(T)\tau}{\rho C_p[1+k(T)\tau]}$$

Visando facilitar a análise dos efeitos das variações dos parâmetros envolvidos no modelo, propõe-se a definição de parâmetros adimensionais que englobem tais efeitos. As formas usuais para esse modelo são: $\beta = \dfrac{UA}{F\rho C_p}$ e $\alpha = \dfrac{(-\Delta H_r)C_{Ae}}{\rho C_p T_e}$, resultando em:

$$\frac{T - T_e}{T_e} + \beta\frac{T - T_w}{T_e} = \alpha\frac{k(T)\tau}{1+k(T)\tau}$$

Definindo agora as variáveis adimensionais: $\theta = \dfrac{T}{T_e}$ e $\theta_w = \dfrac{T}{T_e}$, tem-se:

$$\frac{T - T_e}{T_e} + \beta\frac{T - T_w}{T_e} = \theta - 1 + \beta(\theta - \theta_w)$$

Considerando: $k(T) = k(T_e)\dfrac{k(T)}{k(T_e)}$ e definindo $k_e = k(T_e)$ e $\gamma = \dfrac{E}{RT_e}$, tem-se

$$k(T) = k_e \exp\left[\gamma\left(1 - \frac{T_e}{T}\right)\right] = k_e \exp\left(\gamma\frac{\theta - 1}{\theta}\right).$$

Definindo, finalmente, o parâmetro adimensional: $Da = k_e\tau$ conhecido como o *número da Damköhler*[2], chega-se à equação não linear:

$$f(\theta) = \theta - 1 + \beta(\theta - \theta_w) - \alpha\frac{Da\exp\left(\gamma\dfrac{\theta - 1}{\theta}\right)}{1 + Da\exp\left(\gamma\dfrac{\theta - 1}{\theta}\right)} = 0.$$

[2]Gerhardt Damköhler (1908-1944).

A concentração de saída do reator é definida em termos da variável adimensional:

$$x(\theta) = 1 - X(\theta) = \frac{C_A}{C_{Ae}} = \frac{1}{1 + Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}, \text{ sendo } X(\theta) \text{ a conversão.}$$

Na Figura 4.2 plotam-se as curvas $f(\theta)$ *versus* θ, utilizando os seguintes valores dos parâmetros: $\alpha = \beta = 1$, $\gamma = 20$, $Da = 0,1$ e três valores de $\theta_w \{1,0; 0,8; 0,6\}$.

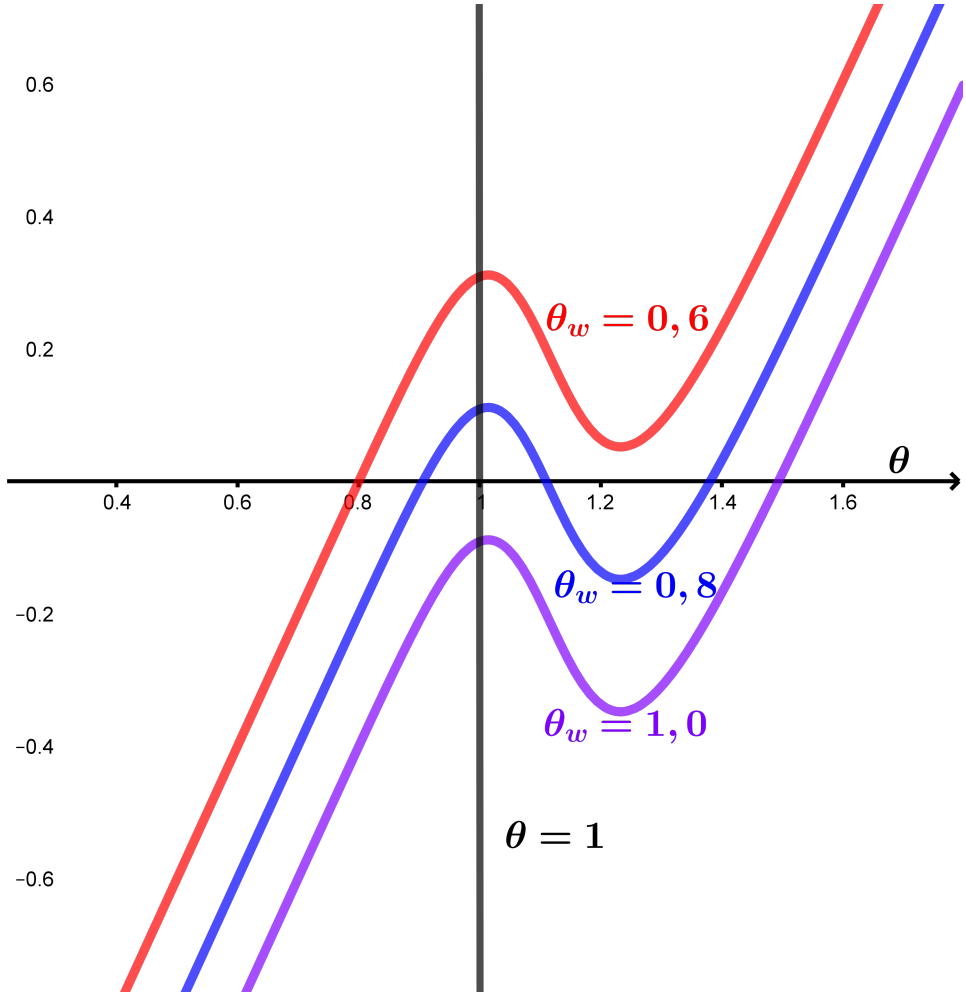


Figura 4.2: Balanço estacionário de energia do CSTR (em forma adimensional).

As raízes de $f(\theta)$ podem ser interpretadas como a interseção da função *calor retirado*: $q_r(\theta) = \theta - 1 + \beta(\theta - \theta_w)$ com a função *calor gerado*: $q_g(\theta) = \alpha \dfrac{Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}{1 + Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}$. Esta forma de representação é mostrada nas várias curvas da Figura 4.3.

Os recursos gráficos e computacionais disponíveis atualmente simplificaram bastante a busca das raízes de equações não lineares, permitindo uma localização preliminar das mesmas apenas pela visualização dos gráficos das funções. Além disto, há grande número de *solvers* que determinam de forma precisa as raízes desejadas, bastando apenas especificar seus valores aproximados ou o intervalo em que se encontram.Tais códigos computacionais contêm procedimentos numéricos

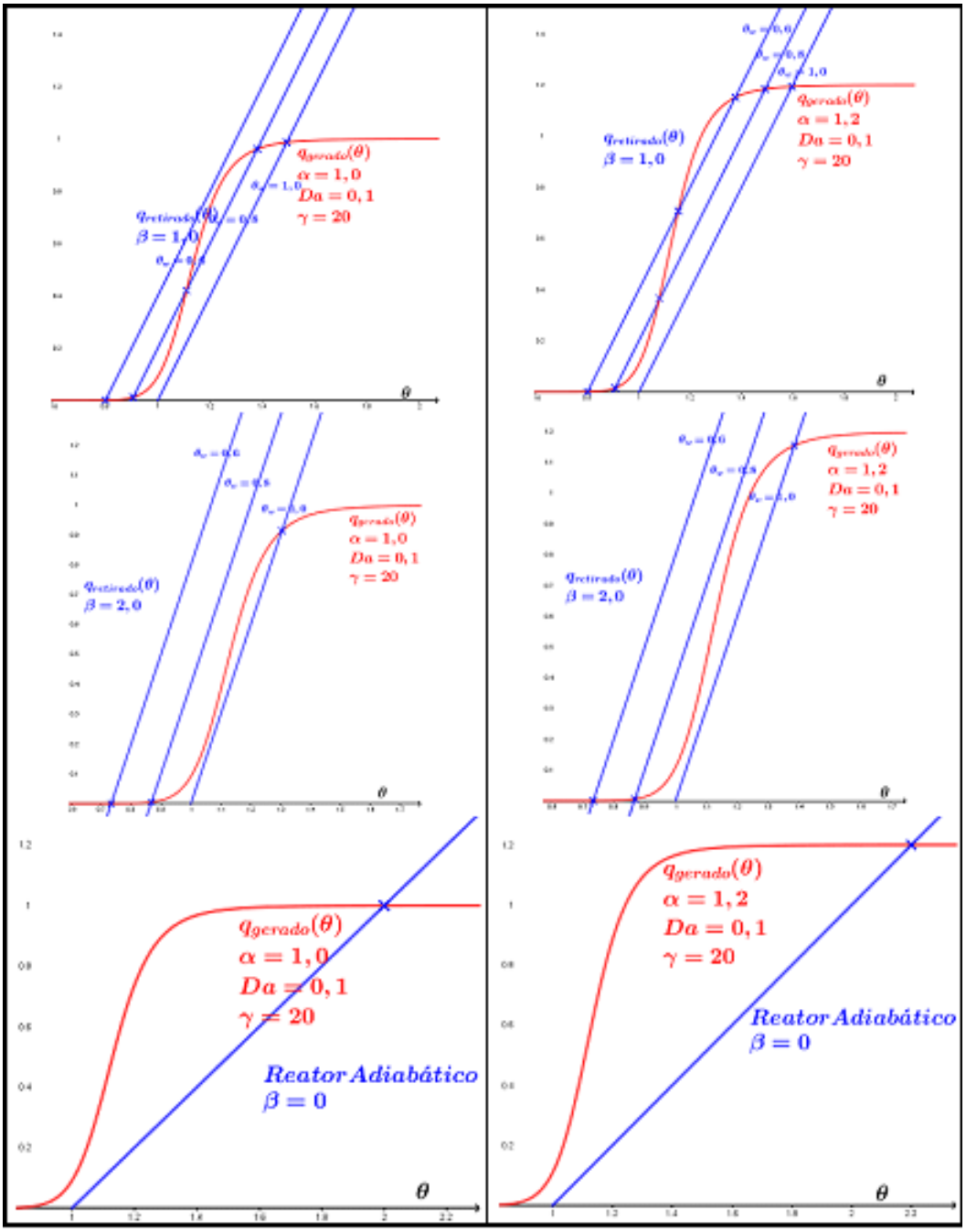


Figura 4.3: Curvas do calor retirado e do calor gerado.

de resolução de equações não lineares de $f(x) = 0, x \in \Re$, a descrição de tais procedimentos é o objetivo deste capítulo. Os métodos mais comumente empregados são:

- Substituições Sucessivas (substituição direta, iteração de ponto fixo ou tentativa-e-erro);
- Newton ou Newton–Raphson[3];
- Newton Modificado;
- Newton Secante ou simplesmente Secante;
- *Regula Falsi* ou Falsa Posição;

[3]Joseph Raphson (1648-1715).

- *Regula Falsi* modificado;
- Bisseção ou Busca Dicotômica;
- Busca Aleatória.

Os métodos que não utilizam as derivadas da função são chamados de *métodos diretos*, sendo considerado como tais o Método da Bisseção e o Método da Busca Aleatória, apresentados na parte inicial deste capítulo. Os Métodos da Secante, *Regula Falsi* e *Regula Falsi* modificado podem também ser considerados como métodos diretos, entretanto, neste capítulo são apresentados como formas distintas de *Métodos Quasi-Newton*, discutidos na Seção 4.7.

4.2 Métodos Diretos

Todos os métodos diretos de busca de raízes de equações algébricas não lineares em uma variável iniciam com a busca do intervalo em que, obrigatoriamente, a raiz deve estar contida. Sendo a a extremidade inferior do intervalo e b a extremidade superior do intervalo, para que exista necessariamente *pelo menos* uma raiz no intervalo deve-se ter:

$$f(a)\, f(b) < 0.$$

No procedimento recursivo de busca da raiz, para que o próximo ponto esteja contido entre a e b se estabelece que:

$$x = a + \lambda\,(b-a), \text{ sendo necessariamente } 0 \leq \lambda \leq 1.$$

Os métodos diretos diferem entre si apenas pela forma com que o parâmetro λ é gerado em cada iteração.

4.2.1 Método da Bisseção

No método da bisseção, o valor de λ é mantido constante e igual a $\frac{1}{2}$ durante todo o processo de busca. O procedimento recursivo pode ser sumarizado por:

$$x^{(k+1)} = \frac{x_L^{(k)} + x_R^{(k)}}{2} \text{ para } k = 0, 1, 2, \cdots$$

sendo $x_L^{(k)}$ o valor da variável x que está à esquerda da raiz na iteração k, $x_R^{(k)}$ o valor da variável x que está à direita da raiz na iteração k, e $x^{(k+1)}$ o valor da variável x na próxima iteração. Após a geração de $x^{(k+1)}$, o novo ponto deve substituir a extremidade que apresenta o mesmo sinal da função neste ponto, assim:

$$\left[x_L^{(k+1)},\, x_R^{(k+1)}\right] = \begin{cases} \left[x^{(k+1)},\, x_R^{(k)}\right] & \text{se } f(x^{(k+1)})\, f(x_L^{(k)}) > 0 \\ \left[x_L^{(k)},\, x^{(k+1)}\right] & \text{se } f(x^{(k+1)})\, f(x_L^{(k)}) < 0 \end{cases}.$$

O processo recursivo é inicializado com $x_L^{(0)} = a$ e $x_R^{(0)} = b$ sendo os valores de a e b obtidos após uma busca preliminar em um intervalo de busca pré-estabelecido tal que $f(a)\, f(b) < 0$. A Figura 4.4 ilustra esse processo iterativo.

Como ao fim de cada iteração o intervalo de busca da raiz é reduzido à metade do valor anterior, pode-se inferir o valor do intervalo na iteração k por: $\Delta^{(k)} = \dfrac{b-a}{2^k}$, deste modo se o procedimento iterativo terminar quando $\Delta^{(n)} = \dfrac{b-a}{2^n} < \varepsilon$, é possível estimar o número máximo de iterações por:

$$k_{maximo} = \text{ceil}\left[\log_2\left(\frac{b-a}{\varepsilon}\right)\right].$$

Para ilustrar esse cálculo considera-se $(b-a) = 10$ e diferentes valores de ε conforme mostrado a seguir:

ε	10^{-6}	10^{-9}	10^{-12}
Número máximo de iterações	24	34	44

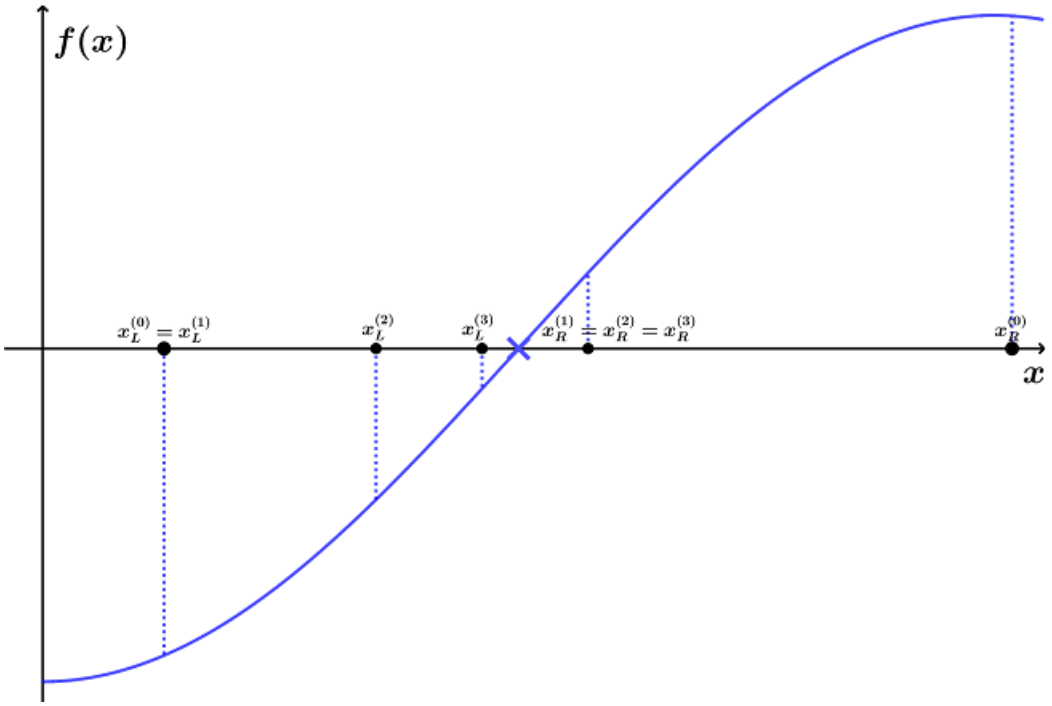


Figura 4.4: Método da bisseção.

4.2.2 Método de Busca Aleatória

No método de busca aleatória, utiliza-se um gerador de números aleatórios, específico do equipamento de cálculo que está sendo utilizado, para determinar o valor de λ. Por exemplo, na tabela abaixo são mostrados os 10 primeiros números aleatórios gerados no MATHCAD® no intervalo [0, 1].

sorteio	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
λ	0,989	0,119	0,009	0,532	0,602	0,166	0,451	0,057	0,783	0,520

A programação dos métodos diretos, inclusive aqueles descritos na Seção 4.7 com redução de intervalo de busca de raízes, em uma forma algorítmica de implementá-los, é a seguir apresentada. A diferença entre os métodos está na definição da função $F(a,b)$ que assume as seguintes formas para os métodos da bisseção e busca aleatória, respectivamente:

1) Método da bisseção: $F(a,b) = \frac{1}{2}$

2) Método da busca aleatória: $F(a,b) = \text{rnd}(1)$, em que $\text{rnd}(\alpha)$ é uma função geradora de números aleatórios entre 0 e α.

$f_a \longleftarrow f(a)$
$f_b \longleftarrow f(b)$
Se $f_a\, f_b > 0$ então busque novos valores de a e b
$k \longleftarrow 0$

Faça
$\lambda \longleftarrow F(a,b)$
$x \longleftarrow a + \lambda(b-a)$
$y \longleftarrow f(x)$
Se $y\, f_a > 0$ faça:
$a \longleftarrow x$
$f_a \longleftarrow y$
senão
$b \longleftarrow x$
$f_b \longleftarrow y$
$\Delta \longleftarrow |b-a|$
$k \longleftarrow k+1$
enquanto $(\Delta > \varepsilon$ ou $|y| > \delta)$ e $k < k_{max}$

Ao final do algoritmo, se $k < k_{max}$ então x contém a raiz encontrada de $f(x)$ e y contém o valor de $f(x)$. Se o número máximo de iterações for atingido sem ocorrer a convergência, deve-se modificar o intervalo inicial $[a,b]$ ou aumentar o número máximo de iterações, ou ainda utilizar outro método de cálculo de λ.

4.3 Método das Substituições Sucessivas

No método das substituições sucessivas (também chamado de método do ponto fixo), o processo iterativo é aplicado a um rearranjo da equação algébrica original $f(x) = 0$ que resulte em: $x = g(x)$. Sugerindo um processo iterativo:

$$x^{(k+1)} = g(x^{(k)}) \text{ para } k = 0, 1, 2, \cdots,$$

cuja representação gráfica é esquematizada na Figura 4.5.

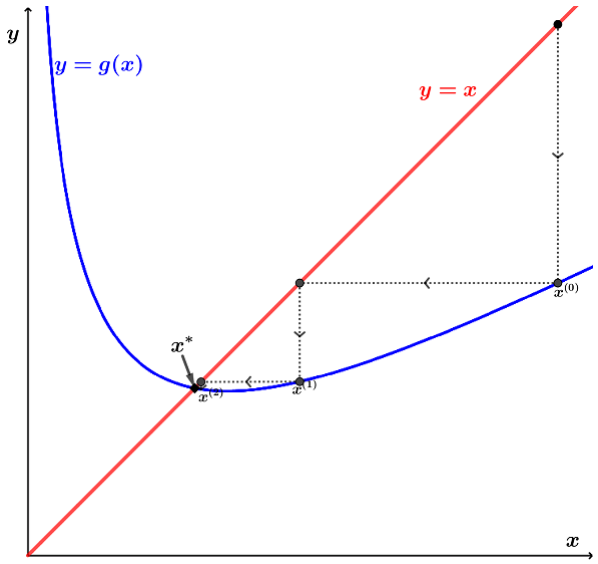


Figura 4.5: Método das substituições sucessivas.

A convergência deste procedimento recursivo para a solução x^* é assegurada se para alguma constante $0 \leq \rho \leq 1$ ocorrer:

$$\left|g(x^{(k)}) - g(x^*)\right| \leq \rho \left|x^{(k)} - x^*\right|,$$

isto é, se $g(x)$ for um **mapeamento contrativo**.

Tal relação pode ser deduzida através da expansão em série de Taylor da função $g(x)$ em torno da solução x^* e truncando a expansão após o segundo termo, assim:

$g(x) \approx g(x^*) + g'(x^*)(x - x^*)$, ou seja: $g(x) - g(x^*) \approx g'(x^*)(x - x^*)$.

Aplicando o módulo a ambos os termos da expressão: $|g(x) - g(x^*)| \leq |g'(x^*)|\ |x - x^*|$.

Comparando esta última expressão com $\left|g(x^{(k)}) - g(x^*)\right| \leq \rho \left|x^{(k)} - x^*\right|$ permite identificar: $|g'(x^*)| \leq \rho < 1$, portanto para que o processo iterativo seja convergente deve-se ter:

$$|g'(x^*)| < 1.$$

Além disto, tendo em vista que: $x^{(k+1)} = g(x^{(k)})$ e $x^* = g(x^*)$ resulta:

$$\left|x^{(k+1)} - x^*\right| \approx |g'(x^*)| \left|x^{(k)} - x^*\right|,$$

caracterizando o método de substituições sucessivas como um método de **convergência linear**.

O gráfico da Figura 4.6 ilustra uma situação em que $|g'(x^*)| > 1$ e $g'(x^*) < 0$, a primeira condição leva a não convergência e a segunda a uma sequência oscilatória em torno da solução.

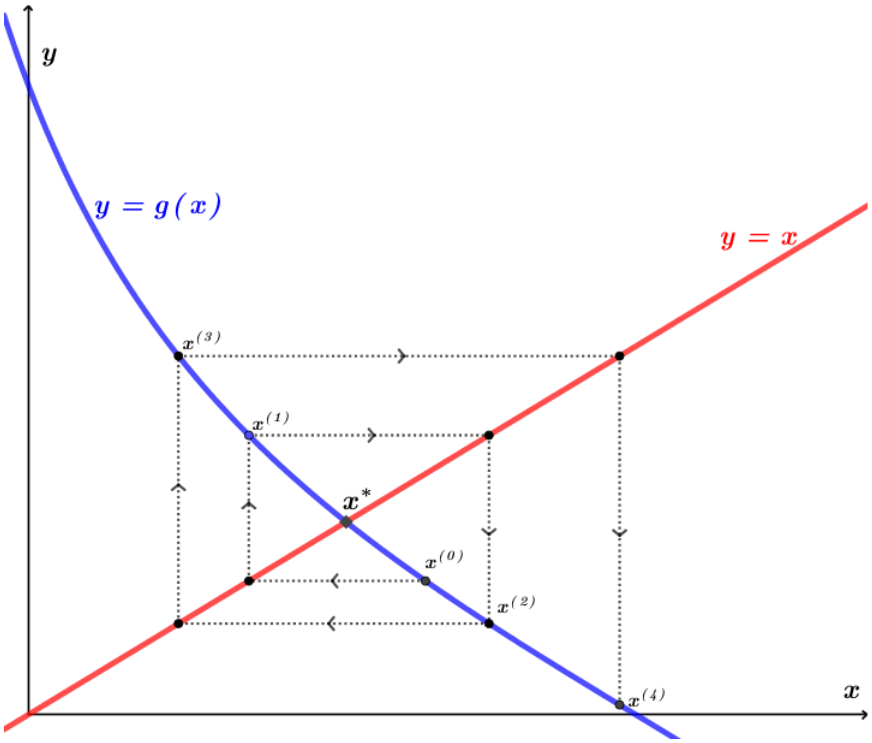


Figura 4.6: Método das substituições sucessivas não convergente e oscilatório.

■ **Exemplo 4.1** Aplicação do método das substituições sucessivas ao modelo estacionário do CSTR.

$$f(\theta) = \theta - 1 + \beta(\theta - \theta_w) - \alpha \frac{Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}{1 + Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)} = 0.$$

A equação é reescrita na forma:

$$\theta = \frac{1+\theta_w}{1+\beta} + \frac{\alpha}{1+\beta}\frac{Da\,\exp\left(\gamma\dfrac{\theta-1}{\theta}\right)}{1+Da\,\exp\left(\gamma\dfrac{\theta-1}{\theta}\right)} = g(\theta)$$

e

$$\frac{dg(\theta)}{d\theta} = \frac{\alpha\gamma}{(1+\beta)\theta^2}\frac{Da\,\exp\left(\gamma\dfrac{\theta-1}{\theta}\right)}{\left[1+Da\,\exp\left(\gamma\dfrac{\theta-1}{\theta}\right)\right]^2}$$

A aplicação do procedimento com os mesmos valores dos parâmetros anteriormente especificados e com $\theta_w = 0,8$ é ilustrada na Figura 4.7 (sem indicação dos eixos!), adotando as condições iniciais: $\theta^{(0)} = 1,08$ e $1,15$.

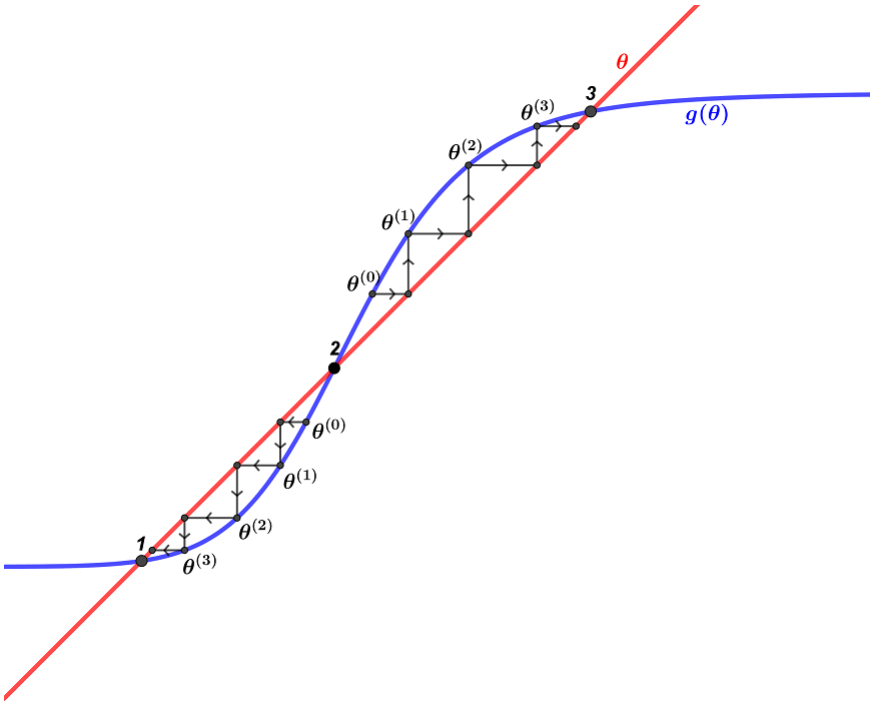


Figura 4.7: Método das Substituições Sucessivas aplicado ao CSTR estacionário.

Os resultados obtidos com as duas condições iniciais são listados a seguir.

k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\theta^{(k)}$	1,08	1,0528	1,0071	0,9516	0,9175	0,9081	0,9065	0,9063	0,9062	0,9062
k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\theta^{(k)}$	1,15	1,1880	1,2515	1,3239	1,3651	1,3773	1,3800	1,3805	1,3806	1,3806

Analisando a Figura 4.7, depreende-se que o problema, apesar de apresentar três soluções distintas, apenas duas destas podem ser obtidas pela aplicação do presente método.

Solução	1	2	3
$\theta^{(*)}$	0,9062	1,1097	1,3806
$\left.\frac{dg(\theta)}{d\theta}\right\|_{\theta^{(*)}}$	-0,1499	-1,9774	-0,1954

A não convergência à solução **2** deve-se ao fato de neste ponto o valor do módulo de $\left.\frac{dg(\theta)}{d\theta}\right|_{\theta^{(*)}}$ ser maior do que 1. Para obter esta solução o método da bisseção poderia ser aplicado, adotando $a = 1,08$ e $b = 1,15$, os resultados da aplicação deste procedimento são apresentados na tabela a seguir.

k	0	1	2	3	4	5	6	7	8	9
$\theta_L^{(k)}$	1,08	1,08	1,0975	1,1063	1,1063	1,1084	1,1095	1,1095	1,1095	1,1097
$\theta_R^{(k)}$	1,15	1,115	1,115	1,115	1,1106	1,1106	1,1106	1,1101	1,1098	1,1098
$\theta_{medio}^{(k)}$	1,115	1,0975	1,1063	1,1106	1,1084	1,1095	1,1101	1,1098	1,1097	1,1097
$f(\theta_{medio}^{(k)})$	-0,0103	0,0235	0,0068	-0,0018	0,0025	0,0004	-0,0007	-0,0001	0,0001	0,0000

Esses resultados mostram que, apesar da simplicidade do método da bisseção e da necessidade de um número elevado de iterações, esse método não apresenta problemas de convergência desde que os valores de a e b sejam criteriosamente selecionados. ■

■ **Exemplo 4.2** Este exemplo visa ilustrar a subjetividade da seleção da função iteração do método das substituições sucessivas considerando a determinação das raízes de $f(x) = e^{-x} - 2\,\text{sen}(x)$. Uma avaliação preliminar das duas funções envolvidas na equação permite concluir que há uma raiz entre 0 e $\frac{\pi}{6}$ pois em $x = 0, e^{-x} = 1$ e $\text{sen}(x) = 0 \Rightarrow f(0) = 1$ já em $x = \frac{\pi}{6}, e^{-x} < 1$ e $2\,\text{sen}(x) = 1 \Rightarrow f(\frac{\pi}{6}) < 0$.

(a) Primeira Seleção: $g(x) = -\ln[2\,\text{sen}(x)]$ neste caso $\frac{dg(x)}{dx} = -\frac{\cos(x)}{\text{sen}(x)} = -\text{cotg}(x) < -1$ no intervalo $0 \leq x \leq \frac{\pi}{6}$, já sendo possível antever a não convergência do método das substituições sucessivas para qualquer valor inicial neste intervalo. O que é confirmado pelos resultados numéricos obtidos e apresentados a seguir, em que se utilizou como valor inicial $x^{(0)} = 0,4$ Verifica-se a não convergência do procedimento, que conduz na quarta iteração a argumento não válido da função logarítmica (na realidade o valor de $x^{(4)}$ é um valor complexo). A representação gráfica do processo é mostrada na Figura 4.8.

k	0	1	2	3	4
$x^{(k)}$	0,4	0,25	0,7038	-0,2579	$0.6732 + 3.1416\boldsymbol{i}$

(b) Segunda Seleção: $g(x) = \text{arcsen}\left(\frac{e^{-x}}{2}\right)$, neste caso $\frac{dg(x)}{dx} = -\frac{e^{-x}}{\sqrt{4 - e^{-2x}}}$, verificando-se que para qualquer valor positivo de x tem-se $\left.\frac{dg(x)}{dx}\right|_{x=0} = -\frac{\sqrt{3}}{3} < \frac{dg(x)}{dx} < 0$, condição que se assegura a convergência do procedimento para qualquer valor inicial positivo de x. O que é confirmado pelos resultados numéricos obtidos e apresentados a seguir, em que se utilizou como valor inicial $x^{(0)} = 0,4$, convergindo à solução, de forma oscilatória, após 8 iterações. A representação gráfica do processo é mostrada na Figura 4.9.

k	0	1	2	3	4	5	6	7	8
$x^{(k)}$	0,4	0,3418	0,3632	0,3551	0,3581	0,3570	0,3574	0,3573	0,3573

■

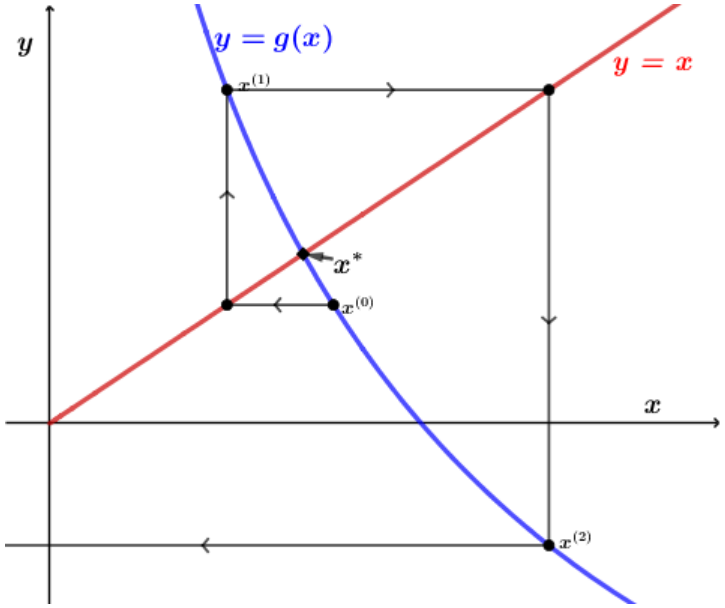


Figura 4.8: Método das substituições sucessivas - Exemplo 4.2 - Primeira Seleção.

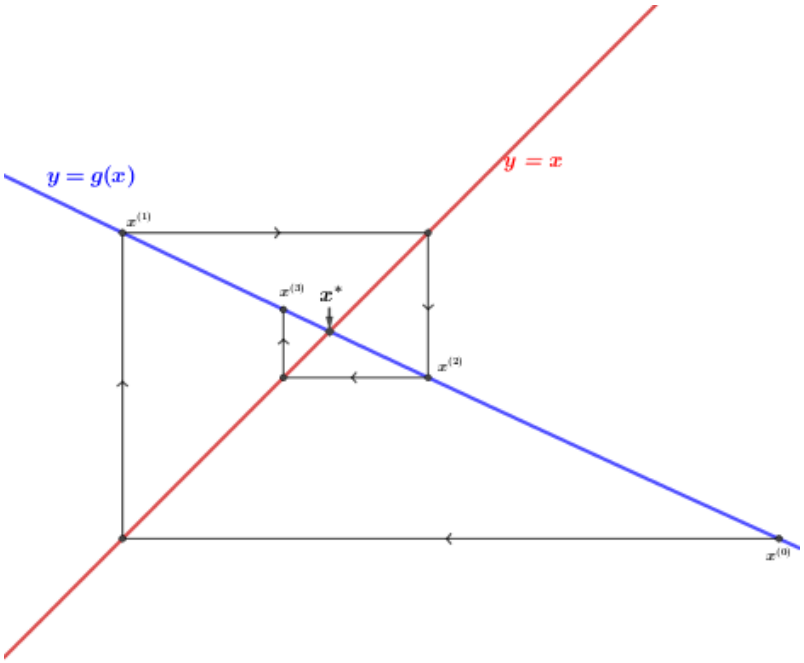


Figura 4.9: Método das substituições sucessivas - Exemplo 4.2 - Segunda Seleção.

4.4 Método de Newton-Raphson

Aproximando a função $f(x)$ por série de potências em torno do valor $x^{(k)}$ tem-se:

$$f(x) \approx f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + \frac{f''(x^{(k)})}{2}(x - x^{(k)})^2 + \cdots$$

Considerando apenas o termo de primeira ordem da aproximação:

$$f(x) \approx f_{linear}(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$$

que é o chamado processo de *linearização* da função.

Considerando agora $x^{(k+1)}$, que é o valor de x na próxima iteração $(k+1)$, como sendo o valor que anula a aproximação linear da função, tem-se:

$$f_{linear}(x^{(k+1)}) = 0 = f(x^{(k)}) + f'(x^{(k)})(x^{(k+1)} - x^{(k)}),$$

resulta:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \quad \text{para} \quad k = 0, 1, 2, \cdots$$

Este procedimento recursivo é o consagrado **Método de Newton-Raphson**, cuja representação geométrica de três iterações sucessivas é esboçada na Figura 4.10. A análise desta figura mostra que o valor de $x^{(k+1)}$ é determinado pela interseção da tangente à curva $f(x)$ no ponto $x^{(k)}$ com o eixo x, e assim sucessivamente.

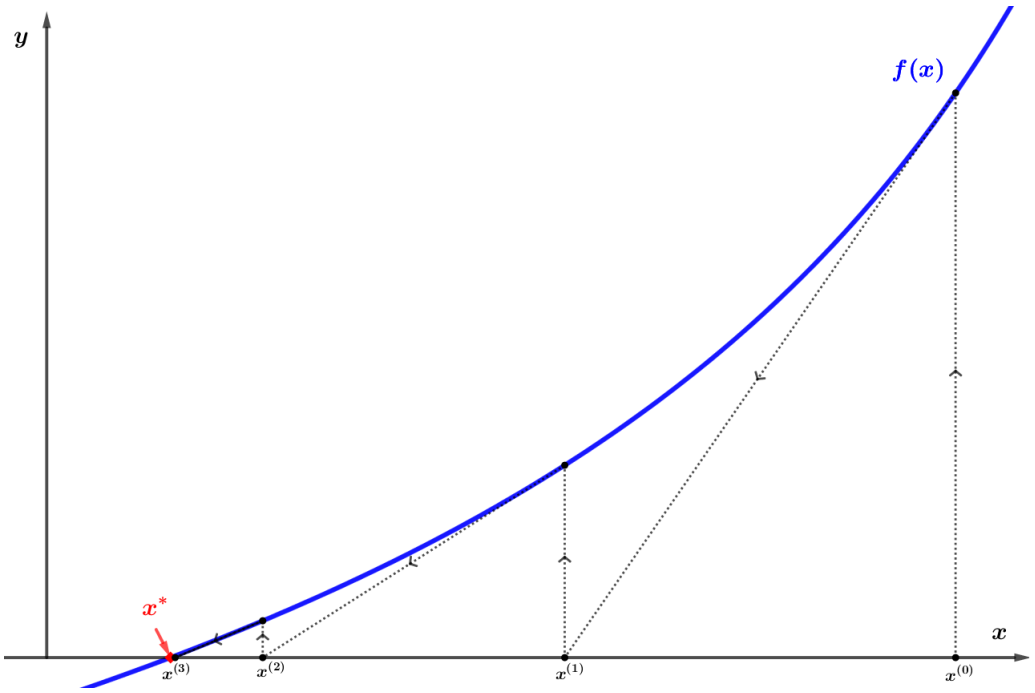


Figura 4.10: Método de Newton-Raphson.

■ **Exemplo 4.3** Com o intuito de demonstrar a eficiência do Método de Newton-Raphson os Exemplos 4.1 e 4.2 são refeitos com a aplicação deste método.

(a) Exemplo 4.1 refeito:

$$\begin{aligned}
f(\theta) &= \theta - 1 + \beta(\theta - \theta_w) - \alpha \frac{Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}{1 + Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)} \\
f'(\theta) &= (1+\beta) + \frac{\alpha\gamma}{\theta^2} \frac{Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)}{\left[1 + Da \exp\left(\gamma \dfrac{\theta - 1}{\theta}\right)\right]^2} \\
\theta^{(k+1)} &= \theta^{(k)} - \frac{f(\theta^{(k)})}{f'(\theta^{(k)})}
\end{aligned}$$

Utilizando as mesmas condições iniciais do Exemplo 4.1, os seguintes resultados são reportados:

k	**0**	**1**	**2**	**3**	**4**
$\theta^{(k)}$	1,08	1,11325	1,10974	1,10973	1,10973
k	**0**	**1**	**2**	**3**	**4**
$\theta^{(k)}$	1,15	1,10516	1,10976	1,10973	1,10973

Verificando-se a convergência para um número menor de iterações e para a solução não obtida pelo método das substituições sucessivas. As outras duas raízes podem ser obtidas partindo de estimativas iniciais mais próximas das respectivas soluções.

(b) Exemplo 4.2 refeito:

$$\begin{aligned} f(x) &= e^{-x} - 2\,\mathrm{sen}(x) \\ f'(x) &= -e^{-x} - 2\cos(x) \\ x^{(k+1)} &= x^{(k)} + \frac{e^{-x^{(k)}} - 2\,\mathrm{sen}(x^{(k)})}{e^{-x^{(k)}} + 2\cos(x^{(k)})} \end{aligned}$$

k	**0**	**1**	**2**	**3**
$x^{(k)}$	0,4	0,35681	0,35733	0,35733

Verificando-se novamente a convergência para um número menor de iterações do que na solução da segunda seleção do método das substituições sucessivas e para a solução não obtida pela primeira seleção.

■

O método de Newton-Raphson pode também ser interpretado como um método de substituições sucessivas em que $g(x) = x - \dfrac{f(x)}{f'(x)}$, deste modo a análise de convergência é feita de maneira similar ao daquele método. Assim, a convergência do método de Newton-Raphson pode ser deduzida através da expansão em série de Taylor dessa forma da função $g(x)$ em torno da solução x^*, assim:

$g(x) \approx g(x^*) + g'(x^*)(x - x^*) + \dfrac{g''(x^*)}{2}(x - x^*)^2$. Em vista de: $g(x^*) = x^*$ e

$g'(x) = 1 - \dfrac{f'(x)}{f'(x)} + \dfrac{f(x)f''(x)}{[f'(x)]^2} = \dfrac{f(x)f''(x)}{[f'(x)]^2} \Rightarrow g'(x^*) = 0,$

o que não permite truncar a série de Taylor após o termo de primeira ordem que é inexistente. Assim: $g(x) \approx x^* + \dfrac{g''(x^*)}{2}(x - x^*)^2$. Calculando esta aproximação para o valor de x na iteração k resulta:

$g(x^{(k)}) = x^{(k+1)} \approx x^* + \dfrac{g''(x^*)}{2}(x^{(k)} - x^*)^2 \Rightarrow x^{(k+1)} - x^* \approx \dfrac{g''(x^*)}{2}(x^{(k)} - x^*)^2.$

Sendo: $g''(x^*) = \dfrac{f''(x^*)}{f'(x^*)}$. Assim, aplicando o módulo a ambos os termos da expressão:

$$\left| x^{(k+1)} - x^* \right| \approx \frac{|g''(x^*)|}{2}(x^{(k)} - x^*)^2,$$

caracterizando o método de Newton-Raphson como um método de **convergência quadrática**.

A forma algorítmica do método das substituições sucessivas e do método de Newton-Raphson são análogas, diferindo apenas na seleção da função iteração $g(x)$, o mesmo podendo ser dito sobre o método de Newton-Raphson com derivada numérica (via perturbação ε). Resumindo:

$$g(x) = \begin{cases} \text{Especificada pelo usuário;} \\ x - \dfrac{f(x)}{f'(x)} \text{ no Método de Newton-Raphson com derivada analítica;} \\ x - \dfrac{f(x)}{\left(\dfrac{f(x+\varepsilon)-f(x)}{\varepsilon}\right)} \text{ no Método de Newton-Raphson com derivada numérica.} \end{cases}$$

Especificação dos valores de ε, δ, k_{maximo} e x_0
$k \longleftarrow 0$

Faça
$x_1 \longleftarrow g(x_0)$
$y \longleftarrow f(x_1)$
$\Delta \longleftarrow |x_1 - x_0|$
$x_0 \longleftarrow x_1$
$k \longleftarrow k+1$
enquanto $(\Delta > \varepsilon$ ou $|y| > \delta)$ e $k < k_{maximo}$

Em certas situações o método de Newton-Raphson tem comportamento inadequado, principalmente em regiões em que ocorrem mudanças da concavidade das funções ou em pontos próximos a pontos em que ocorrem derivadas nulas (pontos de máximo ou mínimo locais ou pontos de inflexão), um exemplo deste comportamento adverso é mostrado na Figura 4.11.

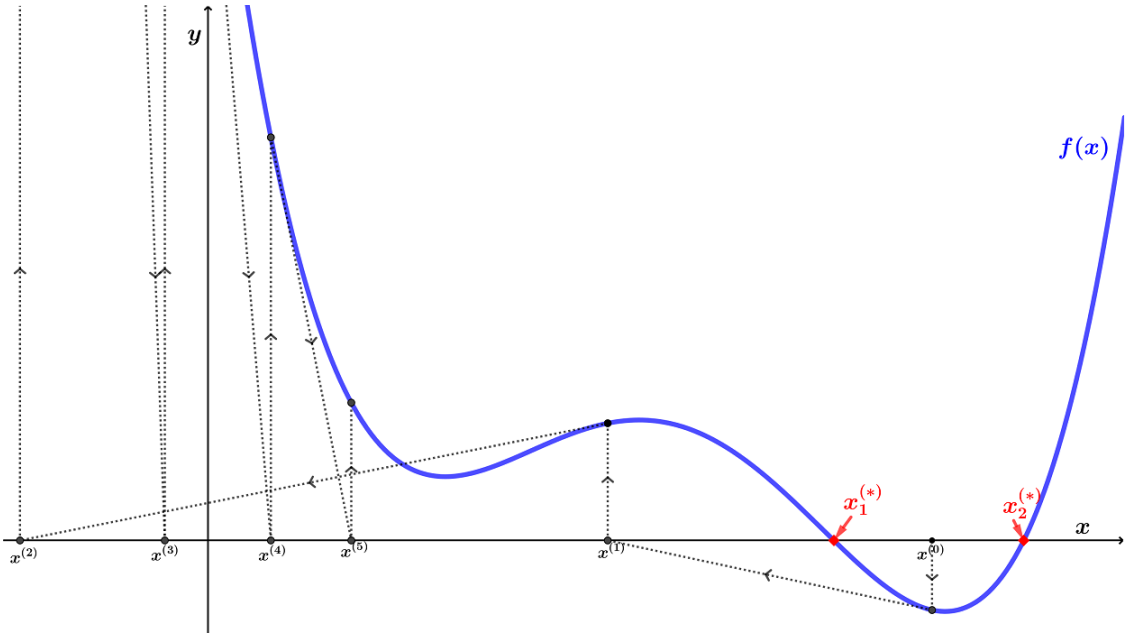


Figura 4.11: Método de Newton-Raphson - caso desfavorável.

Uma familiarização gráfica com a função que se deseja buscar a raiz ou raízes é essencial para assegurar a convergência do método, certificando-se que $x^{(0)}$ esteja contido dentro de um intervalo em que há apenas uma raiz. Condições suficientes de convergência para o método de Newton-Raphson são estabelecidas pelo seguinte teorema.

Teorema 4.4.1 — Condição suficiente de convergência do método de Newton-Raphson. Se $f(x)$ for uma função duas vezes diferenciável no intervalo fechado $[a,b]$ e se as seguintes condições forem satisfeitas:

(i) $f(a)f(b) < 0$;

(ii) $f'(x) \neq 0, x \in [a,b]$;

(iii) $f''(x) \geq 0$ ou $f''(x) \leq 0, \forall x \in [a,b]$;

(iv) Nos pontos extremos a, b deve-se ter $\dfrac{|f(a)|}{|f'(a)|} < (b-a)$ e $\dfrac{|f(b)|}{|f'(b)|} < (b-a)$.

Então o método de Newton-Raphson converge para uma solução única x^* de $f(x) = 0$ em $[a,b]$, para qualquer valor inicial $x^{(0)} \in [a,b]$.

As condições (i) e (ii) garantem que há apenas uma solução em $[a,b]$. A condição (iii) estabelece que não há mudança de concavidade no intervalo $[a,b]$, implicando em $f'(x)$ ser monótona crescente ou decrescente no intervalo. A condição (iv) garante que toda reta tangente à curva no intervalo $[a,b]$, intercepta o eixo x em pontos dentro do mesmo intervalo.

Deve-se destacar que as condições estabelecidas pelo teorema são condições *suficientes*, o que implica em dizer que se as mesmas não forem satisfeitas o processo iterativo ainda assim pode convergir.

É importante ressaltar que o método de Newton-Raphson não se restringe à determinação de raízes reais podendo também ser empregado na determinação de raízes complexas, entretanto, a convergência a raízes complexas só ocorrerá se álgebra complexa for considerada na implementação do método.

4.5 Versões Modificadas do Método de Newton-Raphson

Uma modificação simples no método de Newton é considerar constante a derivada da função $f(x)$ durante todo, ou parte, do processo iterativo, conforme indicado a seguir.

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(0)})} \text{ para } k = 0, 1, 2, \cdots, m-1$$

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(m)})} \text{ para } k = m, (m+1), (m+2), \cdots, 2m-1$$

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(2m)})} \text{ para } k = 2m, (2m+1), (2m+2), \cdots, 3m-1$$

Na Figura 4.12, o método de Newton-Raphson usual é confrontado com o método de Newton-Raphson modificado no qual o valor da derivada da função em $x^{(0)}$ é *congelada*, mantendo-se com este valor ao longo de todo o processo iterativo.

Alega-se que a avaliação da derivada da função em menos pontos reduz o *custo* computacional de cada iteração. Isso pode ser relevante para o caso multivariável de dimensão elevada, porém esta economia pode ser comprometida pelo maior número de iterações para haver convergência à solução.

Modificações de ordens mais elevadas, que convergem mais rapidamente, baseiam-se na aproximação de $f(x)$ por polinômio de Taylor de segundo grau obtido pela expansão em torno do $x^{(k)}$ assim:

$$f(x) \approx p_2(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + \frac{f''(x^{(k)})}{2}(x - x^{(k)})^2.$$

Calculando a seguir $x^{(k+1)}$ que anula $p_2(x)$ ou seja:

$$f(x^{(k)}) + f'(x^{(k)})(x^{(k+1)} - x^{(k)}) + \frac{f''(x^{(k)})}{2}(x^{(k+1)} - x^{(k)})^2 = 0,$$

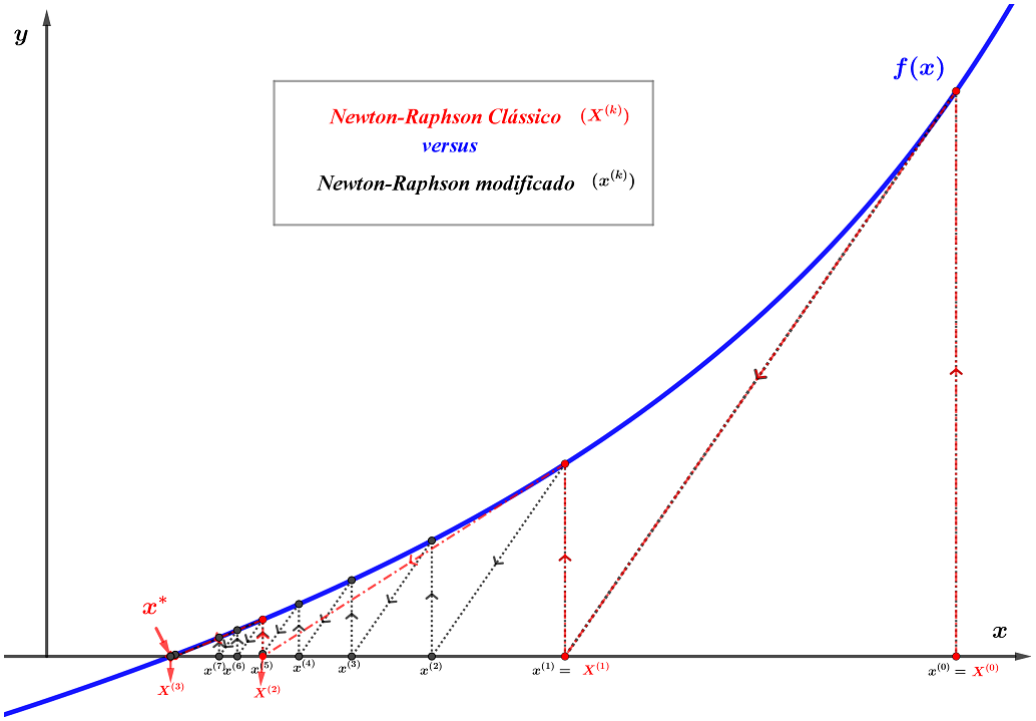


Figura 4.12: Método de Newton-Raphson clássico e modificado.

que pode ser rearranjada na forma:

$$x^{(k+1)} - x^{(k)} = -\frac{f(x^{(k)})}{f'(x^{(k)})}\left[\frac{1}{1+\dfrac{f''(x^{(k)})[x^{(k+1)}-x^{(k)}]}{2f'(x^{(k)})}}\right].$$

O método de Richmond[4] de terceira ordem considera o valor de $(x^{(k+1)} - x^{(k)})$ no termo entre colchetes da equação anterior como o obtido pelo método de Newton-Raphson, isto é, $x^{(k+1)} - x^{(k)} = -\dfrac{f(x^{(k)})}{f'(x^{(k)})}$, resultando finalmente em:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}\left[\frac{1}{1-\dfrac{1}{2}\dfrac{f''(x^{(k)})f(x^{(k)})}{[f'(x^{(k)})]^2}}\right].$$

À medida que o valor de $x^{(k)}$ converge para a solução $f(x^{(k)}) \longrightarrow 0$ o método de Richmond tende à forma usual do método de Newton-Raphson. A forma algorítmica de implementação do método de Richmond é análoga às formas algorítmicas do método das substituições sucessivas e do método de Newton-Raphson, bastando definir a função iteração como:

$$g(x) = x - \frac{f(x)}{f'(x)}\left[\frac{1}{1-\dfrac{1}{2}\dfrac{f''(x)f(x)}{[f'(x)]^2}}\right].$$

[4]Herbert William Richmond (1863-1948).

Um aperfeiçoamento do método de Newton-Raphson pode ser feito baseado na constatação do mesmo ser um método de convergência quadrática, implicando em:

$$x^{(k+1)} - x^* \approx p(x^{(k)} - x^*)^2.$$

Assim, considerando duas iterações sucessivas:

$$\begin{aligned}
\varepsilon_0 &= x^{(k)} - x^* = \varepsilon_1 - (x^{(k+1)} - x^{(k)}) \\
\varepsilon_1 &= x^{(k+1)} - x^* \\
\varepsilon_2 &= x^{(k+2)} - x^* = \varepsilon_1 + (x^{(k+2)} - x^{(k+1)})
\end{aligned}$$

Para simplificar a notação, adotam-se:
$\alpha = (x^{(k+1)} - x^{(k)})$, $\alpha\, q = (x^{(k+2)} - x^{(k+1)})$ e $p = \dfrac{\varepsilon_1}{(\varepsilon_0)^2} = \dfrac{\varepsilon_2}{(\varepsilon_1)^2}$, resultando em:

$$(\varepsilon_1)^3 = (\varepsilon_0)^2\varepsilon_2 = (\varepsilon_1 - \alpha)^2(\varepsilon_1 + \alpha\, q) \Rightarrow (q-2)(\varepsilon_1)^2 + \alpha(1-2q)\varepsilon_1 + q\,\alpha^2 = 0$$

$$\varepsilon_1 = \begin{cases} \dfrac{2\alpha}{3} \text{ se } q = 2; \\ \dfrac{\alpha}{2(q-2)}\left[(2q-1) \pm \sqrt{1+4q}\right] \text{ se } q \neq 2. \end{cases}$$

Assim: $x^* \approx x^{(k+1)} - \varepsilon_1$, selecionando, se $q \neq 2$, o valor de ε_1 no qual $|f(x^{(k+1)} - \varepsilon_1)|$ apresenta o menor valor. A seguir, o valor de $x^{(k+2)}$ é substituído por $x^{(k+1)} - \varepsilon_1$ e os dois valores seguintes, $x^{(k+3)}$ e $x^{(k+4)}$, são calculados pelo método de Newton-Raphson convencional, o valor de $x^{(k+4)}$ é aprimorado da mesma forma aplicada ao valor de $x^{(k+2)}$ e assim sucessivamente.

A comparação do desempenho desta forma *aprimorada* do método de Newton-Raphson com o desempenho da forma convencional é ilustrada com a busca da menor raiz real positiva da função $f(x) = \text{tg}(\frac{x}{5}) - 1$, cuja solução é $x^* = 3,9269908$, adotando como condição inicial $x^{(0)} = 7,7$.

k	$x^{(k)}$ Newton-Raphson convencional	$x^{(k)}$ Newton-Raphson *aprimorado*
0	7,7000000	7,7000000
1	7,5508563	7,5508563
2	7,2668276	3,0702203
3	6,7536902	4,0556978
4	5,9268229	3,9274927
5	4,8916624	3,9269909
6	4,1345727	**3,9269908**
7	3,9358424	
8	3,9270065	
9	**3,9269908**	

4.6 Determinação das Raízes de Polinômios de Coeficientes Reais

Quando a função $f(x)$ for um polinômio, geralmente, deseja-se obter todas suas raízes. Neste caso os métodos numéricos podem ser adaptados para esse tipo especial de função, em especial, formas apropriadas para calcular o valor do polinômio e de sua derivada são apresentadas. Além disso, regras práticas para a localização e caracterização da natureza das raízes, bem como procedimentos numéricos que possibilitem a determinação de todas as raízes são apresentados.

1. **Cálculo do valor do polinômio e de sua derivada para um valor genérico do argumento**

   Considerando $f(x) = p_n(x)$ um polinômio de coeficientes reais da forma $p_n(x) = \sum_{i=0}^{n} a_i x^i$.

   A aplicação do método de Newton-Raphson para a determinação das raízes de $p_n(x)$ dá origem ao procedimento iterativo:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \text{ para } k = 0, 1, 2, \cdots$$

   A avaliação do valor do polinômio e de sua derivada para um valor genérico de seu argumento pode ser feita por meio de um procedimento recursivo sugerido por Horner, que evita o cálculo das sucessivas potências do argumento. Este procedimento é baseado no fato de o valor de um polinômio de qualquer grau, para um valor genérico de $x = \alpha$, ser igual ao resto da divisão do polinômio pelo monômio $(x - \alpha)$, uma vez que:

$$\frac{p_n(x)}{x-\alpha} = q_{n-1}(x) + \frac{\rho}{x-\alpha} \Rightarrow p_n(x) = (x-\alpha)q_{n-1}(x) + \rho.$$

   Substituindo x por α nessa última expressão, chega-se a $p_n(\alpha) = \rho$, demonstrando assim o estabelecido por Horner.

   Com a notação indicial: $q_{n-1}(x) = \sum_{i=0}^{n-1} b_{i+1} x^i$ e $\rho = b_0$, implica:

$$(x-\alpha)q_{n-1}(x) + \rho = b_n x^n + \sum_{i=0}^{n-1} (b_i - \alpha b_{i+1}) x^i = \sum_{i=0}^{n} a_i x^i.$$

   Dando origem ao procedimento recursivo:

$$\begin{aligned} b_n &= a_n \\ b_i &= a_i + \alpha b_{i+1} \text{ para } i = n-1, n-2, \cdots, 0. \end{aligned}$$

   Sendo $b_0 = \rho = p_n(\alpha)$. O valor da derivada de $p_n(x)$ em $x = \alpha$, pode ser determinada através da diferenciação de ambos os membros da expressão:

$$p_n(x) = (x-\alpha)q_{n-1}(x) + \rho \Rightarrow p'_n(x) = (x-\alpha)q'_{n-1}(x) + q_{n-1}(x),$$

   com $x = \alpha$, obtém-se: $p'_n(\alpha) = q_{n-1}(\alpha)$. Para calcular $q_{n-1}(\alpha)$, basta repetir o procedimento aplicado na determinação de $p_n(\alpha)$, assim:

$$\frac{q_{n-1}(x)}{x-\alpha} = q_{n-2}(x) + \frac{c_0}{x-\alpha} \Rightarrow q_{n-1}(x) = (x-\alpha)q_{n-2}(x) + c_0.$$

   Representando $q_{n-2}(x) = \sum_{i=0}^{n-2} c_{i+1} x^i$ implica:

$$(x-\alpha)q_{n-2}(x) + c_0 = c_{n-1}x^{n-1} + \sum_{i=0}^{n-2} (c_i - \alpha c_{i+1}) x^i = \sum_{i=0}^{n-1} b_{i+1} x^i.$$

   Dando origem ao procedimento recursivo:

$$\begin{aligned} c_{n-1} &= b_n \\ c_i &= b_{i+1} + \alpha c_{i+1} \text{ para } i = n-2, n-3, \cdots, 0. \end{aligned}$$

   Sendo $c_0 = \rho = p'_n(\alpha)$.

2. **Localização preliminar das raízes**
   (a) **Regra de Sinais de Descartes**[5]
   Se os termos de um polinômio com coeficientes reais são colocados em ordem decrescente de grau, então o número de raízes reais positivas do polinômio é ou igual ao número de permutações de sinal ou menor por uma diferença par.
   Em decorrência desta regra, pode-se também afirmar: Se os termos de um polinômio com coeficientes reais, após a troca de sinal de seu argumento, são colocados em ordem decrescente de grau, então o número de raízes reais negativas do polinômio é ou igual ao número de permutações de sinal ou menor por uma diferença par.
   Para ilustrar essa regra, o seguinte exemplo é apresentado:

$$p_7(x) = 40x^7 - 36x^6 - 70x^5 + 123x^4 + 76x^3 - 177x^2 - 46x + 90$$

$$p_7(-x) = -40x^7 - 36x^6 + 70x^5 + 123x^4 - 76x^3 - 177x^2 + 46x + 90$$

   Número de permutações de sinal em $p_7(x)$ $n = 4$, assim o número de raízes reais positivas é igual a 4, ou 2 ou 0.
   Número de permutações de sinal em $p_7(-x)$ $n = 3$, assim o número de raízes reais negativas é igual a 3, ou 1.
   Baseado nessas assertivas, constrói-se a tabela:

Possibilidade	Número de raízes reais positivas	Número de raízes reais negativas	Número de pares de raízes complexas
1	4	3	0
2	4	1	1
3	2	3	1
4	2	1	2
5	0	3	2
6	0	1	3

   (b) **Estabelecimento do domínio máximo da localização das raízes no plano complexo**

   Todas as raízes de $p_n(x)$ localizam-se, no plano complexo, Figura 4.13, no interior do círculo de raio ρ (*raio espectral*) determinado por:

$$\rho = 2\max\left[\left|\frac{a_i}{a_n}\right|^{\left(\frac{1}{n-i}\right)} \text{ para } i = 0, 1, \cdots, (n-1)\right].$$

[5] René Descartes (1596-1650).

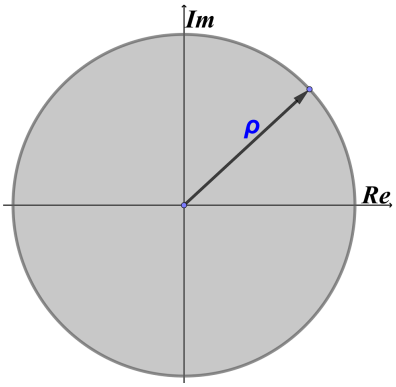


Figura 4.13: Região das raízes no plano complexo.

Aplicação da regra ao exemplo anterior:
$p_7(x) = 40x^7 - 36x^6 - 70x^5 + 123x^4 + 76x^3 - 177x^2 - 46x + 90.$
$\rho = 2\max\left[(\frac{9}{4})^{\frac{1}{7}}, (\frac{23}{20})^{\frac{1}{6}}, (\frac{177}{40})^{\frac{1}{5}}, (\frac{19}{10})^{\frac{1}{4}}, (\frac{123}{40})^{\frac{1}{3}}, (\frac{7}{4})^{\frac{1}{2}}, (\frac{9}{10})\right] = 2,9083.$

3. **Determinação das raízes pelo método de *Newton-Raphson***
Após a representação gráfica do polinômio, pode-se caracterizar a natureza de suas raízes, assim para determinar a menor raiz real aplica-se diretamente o método de Newton-Raphson adotando como estimativa inicial $x^{(0)} = -\rho$ de acordo com o processo iterativo:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \text{ para } k = 0, 1, 2, \cdots \text{ (até a convergência), com } x^{(0)} = -\rho$$

Calculando $p_n(x^{(k)})$ e $p'_n(x^{(k)})$ pelo procedimento descrito no item 1. O valor convergido desse processo é a menor raiz real positiva e será designada por x_1. Para determinar a maior raiz real, repete-se o procedimento adotando como estimativa inicial $x^{(0)} = \rho$, e o valor convergido nesse segundo processo será denominado x_{max}.
Para determinar a segunda raiz real aplica-se o método de Newton-Raphson ao polinômio *deflatado*, que é um polinômio da grau $(n-1)$ resultante da divisão:

$$p_{n-1}(x) = \frac{p_n(x)}{x - x_1}.$$

Aplicando o método de Newton-Raphson ao polinômio deflatado, resulta:

$$x^{(k+1)} = x^{(k)} - \frac{p_{n-1}(x^{(k)})}{p'_{n-1}(x^{(k)})}.$$

Um artifício empregado, que evita a divisão do polinômio original por $(x - x_1)$, é desenvolvido aplicando o logaritmo neperiano a ambos os lados de:
$p_{n-1}(x) = \frac{p_n(x)}{x - x_1}$, ou seja, $\ln(p_{n-1}(x)) = \ln(p_n(x)) - \ln(x - x_1)$ e derivando ambos os lados dessa última expressão, o que resulta em:

$$\frac{p'_{n-1}(x)}{p_{n-1}(x)} = \frac{p'_n(x)}{p_n(x)} - \frac{1}{(x - x_1)} = \frac{p'_n(x)}{p_n(x)}\left[1 - \frac{p_n(x)}{p'_n(x)}\frac{1}{(x - x_1)}\right]$$

Logo: $\frac{p_{n-1}(x)}{p'_{n-1}(x)} = \frac{p_n(x)}{p'_n(x)}\left[\frac{1}{1 - \frac{p_n(x)}{p'_n(x)}\frac{1}{(x - x_1)}}\right]$

O que dá origem ao procedimento:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left[\frac{1}{1 - \dfrac{p_n(x^{(k)})}{p'_n(x^{(k)})} \dfrac{1}{(x^{(k)} - x_1)}} \right] \text{ para } k = 0, 1, 2, \cdots \text{ (até a convergência),}$$

com $x^{(0)} = x_1 + \varepsilon$, sendo ε um valor próximo de zero e positivo.
Esse procedimento além de evitar a divisão do polinômio original por $(x - x_1)$, não acrescenta à imprecisão numérica da segunda raiz a imprecisão numérica da primeira raiz e, a medida que se converge à segunda raiz, o procedimento aproxima-se do método de Newton-Raphson aplicado diretamente ao polinômio original.
A repetição do mesmo artifício pode ser generalizado na determinação da raiz x_m após as raízes reais $x_1 < x_2 < \cdots < x_{m-1}$ serem determinadas:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left[\frac{1}{1 - \dfrac{p_n(x^{(k)})}{p'_n(x^{(k)})} \displaystyle\sum_{j=1}^{m-1} \frac{1}{(x^{(k)} - x_j)}} \right] \text{ para } k = 0, 1, 2, \cdots \text{ (até a convergência),}$$

com $x^{(0)} = x_{m-1} + \varepsilon$, sendo ε um valor próximo de zero e positivo.
Quando a raiz real obtida por esse processo for igual a x_{max} é um indicativo que não há mais raiz real e as restantes são pares conjugados de raízes complexas. Para buscar a primeira raiz complexa basta repetir o processo (usando álgebra complexa):

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left[\frac{1}{1 - \dfrac{p_n(x^{(k)})}{p'_n(x^{(k)})} \displaystyle\sum_{j=1}^{M} \frac{1}{(x^{(k)} - x_j)}} \right] \text{ para } k = 0, 1, 2, \cdots \text{ (até a convergência),}$$

com $x^{(0)} = \rho[\cos(\theta) + \text{sen}(\theta)\boldsymbol{i}]$, condição inicial no plano complexo e situada sobre o círculo delimitante da região das raízes. O índice superior, M, no somatório é o número total de raízes reais.
Seja $x_1^{(*)} = \sigma_1 + \omega_1\boldsymbol{i}$ o valor da raiz complexa que convergiu no último processo iterativo, como as raízes complexas são sempre pares conjugados $x_2^{(*)} = \sigma_1 - \omega_1\boldsymbol{i}$ também será raiz, e a inclusão destas raízes ao termo $\displaystyle\sum_{j=1}^{M} \frac{1}{(x^{(k)} - x_j)}$ é feita com os dois termos:

$$\frac{1}{(x - \sigma_1) - \omega_1\boldsymbol{i}} + \frac{1}{(x - \sigma_1) + \omega_1\boldsymbol{i}} = \frac{2(x - \sigma_1)}{(x - \sigma_1)^2 + \omega_1^2}.$$

Para determinar o segundo par de raízes complexas, o seguinte procedimento iterativo é aplicado a:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left[\frac{1}{1 - \dfrac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left(\displaystyle\sum_{j=1}^{M} \frac{1}{(x^{(k)} - x_j)} + \frac{2(x - \sigma_1)}{(x - \sigma_1)^2 + \omega_1^2} \right)} \right] \text{ para } k = 0, 1, 2, \cdots$$

(até a convergência),

com $x^{(0)} = (\sigma_1 + \varepsilon) + (\omega_1 + \varepsilon)\boldsymbol{i}$, sendo ε um valor próximo de zero e positivo.
Repetindo-se o procedimento para a função iterativa:

$$x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left[\frac{1}{1 - \frac{p_n(x^{(k)})}{p'_n(x^{(k)})} \left(\sum_{j=1}^{M} \frac{1}{(x^{(k)} - x_j)} + 2 \sum_{i=1}^{r-1} \frac{(x-\sigma_i)}{(x-\sigma_i)^2 + \omega_i^2} \right)} \right] \text{ para } k = 0, 1, 2, \cdots \text{ (até a convergência),}$$

com $x^{(0)} = (\sigma_{r-1} + \varepsilon) + (\omega_{r-1} + \varepsilon)\boldsymbol{i}$, sendo ε um valor próximo de zero e positivo. Repetindo-se o procedimento até $M + 2r = n$, o que indica que todas as raízes foram determinadas.

■ **Exemplo 4.4** Determinação ds raízes de:

$$p_8(x) = 16x^8 - 60x^6 + 29x^5 + 88x^4 - \frac{303}{4}x^3 - \frac{243}{2}x^2 + \frac{61}{2}x + 30$$

$\rho = 2\max\left[(\frac{15}{8})^{\frac{1}{8}}, (\frac{61}{32})^{\frac{1}{7}}, (\frac{243}{32})^{\frac{1}{6}}, (\frac{303}{64})^{\frac{1}{5}}, (\frac{11}{2})^{\frac{1}{4}}, (\frac{29}{16})^{\frac{1}{3}}, (\frac{15}{4})^{\frac{1}{2}}, 0\right] = 3,873.$
Explorando a representação gráfica no domínio $-3,9 \leq x \leq +3,9$ chega-se, após um ajuste *fino*, ao domínio $-1,8 \leq x \leq +1,8$, conforme representado na Figura 4.14.

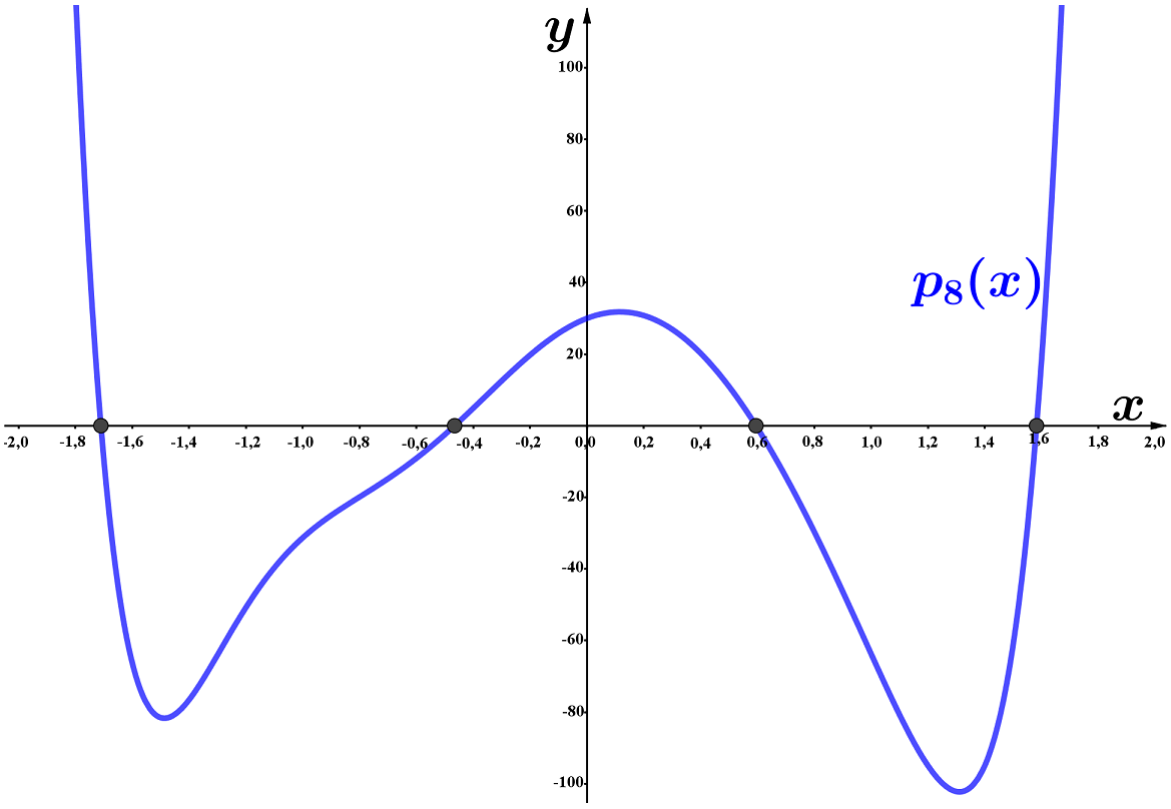


Figura 4.14: Polinômio de oitavo grau.

Os resultados dos procedimentos iterativos aplicados na determinação das oito raízes do polinômio são listados na tabela a seguir.

k	$x^{(k)}$	$x^{(k)}$	$x^{(k)}$	$x^{(k)}$	$x^{(k)}$	$x^{(k)}$
0	-1,8000	-1,7007	-0,4560	1,5920	-0,8491+0,5283i	-0,99+0,51i
1	-1,7337	-1,4563	2,1559	1,2108	-1+0,4679i	0,2408+0,3154i
2	-1,7126	-1,2349	1,9054	0,8645	-0,9991+0,5012i	1,182+0,391i
3	-1,7107	-1,0121	1,7242	0,6393	-1+0,5i	0,6018+1,2465i
4	-1,7107	-0,7442	1,6198	0,5956	-1+0,5i	0,9172+0,9872i
5	-1,7107	-0,4950	1,5854	0,5947	-1+0,5i	1,0008+0,9965i
6	-1,7107	-0,4663	1,5821	0,5947	-1+0,5i	1+1i
7	-1,7107	-0,4660	1,5820	0,5947	-1+0,5i	1+1i

■

A determinação direta das raízes complexas conjugadas pode ser feita pelo método de **Newton-Bairstow**[6]. Nesse procedimento, o método de Newton-Raphson é aplicado de modo que o polinômio original seja fatorável de forma exata pela forma quadrática: $(x-\sigma)^2+\omega^2 = x^2-2\sigma+(\sigma^2+\omega^2) = x^2-p\,x-q$. A vantagem desse procedimento é que determina raízes complexas conjugadas mantendo-se no domínio real.
Essa fatoração é convenientemente expressa na forma:

$$p_n(x) = \sum_{i=0}^{n} a_i x^i = (x^2 - p\,x - q)\sum_{i=0}^{n-2} b_{i+2}x^i + b_1(x-p) + b_0.$$

Dando origem ao procedimento recursivo:

$$\begin{cases} b_n = a_n \\ b_{n-1} = a_{n-1} + p\,b_n \\ b_i = a_i + p\,b_{i+1} + q\cdot b_{i+2} \text{ para } i = n-2, n-3, \ , 1, 0. \end{cases}.$$

A busca iterativa dos valores de p e q que levem a valores nulos de b_0 e b_1 é o objetivo do método. A linearização das expressões de b_0 e b_1 em torno de $\begin{pmatrix} p^* \\ q^* \end{pmatrix}$, resulta em:

$$\begin{cases} b_1(p,q) = 0 \Rightarrow b_1^{(linear)}(p,q) = b_1(p^*,q^*) + \left.\dfrac{\partial b_1}{\partial p}\right|_{(p^*,q^*)}(p-p^*) + \left.\dfrac{\partial b_1}{\partial q}\right|_{(p^*,q^*)}(q-q^*) = 0 \\ b_0(p,q) = 0 \Rightarrow b_0^{(linear)}(p,q) = b_0(p^*,q^*) + \left.\dfrac{\partial b_0}{\partial p}\right|_{(p^*,q^*)}(p-p^*) + \left.\dfrac{\partial b_0}{\partial q}\right|_{(p^*,q^*)}(q-q^*) = 0 \end{cases}.$$

Adotando a notação indicial: $A_{i,0} = \dfrac{\partial b_i}{\partial p}$ e $A_{i,1} = \dfrac{\partial b_i}{\partial q}$, obtém-se do procedimento recursivo de cálculo dos coeficientes b_i:

$$\begin{cases} \dfrac{\partial b_n}{\partial p} = \dfrac{\partial b_n}{\partial q} = 0 \Rightarrow A_{n,0} = A_{n,1} = 0 \\ \dfrac{\partial b_{n-1}}{\partial p} = b_n \text{ e } \dfrac{\partial b_{n-1}}{\partial q} = 0 \Rightarrow A_{n-1,0} = b_n \text{ e } A_{n-1,1} = 0 \\ \dfrac{\partial b_i}{\partial p} = b_{i+1} + p\,\dfrac{\partial b_{i+1}}{\partial p} + q\,\dfrac{\partial b_{i+2}}{\partial p} \\ \dfrac{\partial b_i}{\partial q} = b_{i+2} + p\,\dfrac{\partial b_{i+1}}{\partial q} + q\,\dfrac{\partial b_{i+2}}{\partial q} \\ A_{i,0} = b_{i+1} + p\,A_{i+1,0} + q\,A_{i+2,0} \text{ para } i = n-2, n-3, \cdots, 1, 0 \\ A_{i,1} = b_{i+2} + p\,A_{i+1,1} + q\cdot A_{i+2,1} \text{ para } i = n-2, n-3, \cdots, 1, 0. \end{cases}.$$

[6] Sir Leonard Bairstow (1880-1963).

Utilizando $\begin{pmatrix} p^* \\ q^* \end{pmatrix} = \begin{pmatrix} p^{(k)} \\ q^{(k)} \end{pmatrix}$ os valores de p e q na iteração k, calcula-se $\begin{pmatrix} p^{(k+1)} \\ q^{(k+1)} \end{pmatrix}$ de modo que: $b_0^{(linear)}(p^{(k+1)}, q^{(k+1)}) = b_1^{(linear)}(p^{(k+1)}, q^{(k+1)}) = 0$, assim:

$$\begin{pmatrix} b_1 \\ b_0 \end{pmatrix}_{(p^{(k)}, q^{(k)})} + \begin{pmatrix} A_{1,0} & A_{1,1} \\ A_{0,0} & A_{0,1} \end{pmatrix}_{(p^{(k)}, q^{(k)})} \begin{pmatrix} p^{(k+1)} - p^{(k)} \\ q^{(k+1)} - q^{(k)} \end{pmatrix} = 0.$$

A resolução desse sistema linear em cada iteração permite calcular novos valores de p e q e assim sucessivamente.
O procedimento é aplicado na determinação dos dois pares de raízes complexas do Exemplo 4.4 estando listados os resultados na tabela a seguir.

k	$p^{(k)}$	$q^{(k)}$	$p^{(k)}$	$q^{(k)}$
0	4,0000	-8,0000	-4,0000	-8,0000
1	3,5990	-6,0929	-3,6101	-5,9974
2	3,2373	-4,6734	-3,2641	-4,5213
3	2,9098	-3,6114	-2,9593	-3,4240
4	2,6075	-2,8183	-2,6934	-2,6110
5	2,3158	-2,2356	-2,4617	-2,0169
6	2,0302	-1,8659	-2,2577	-1,5964
7	1,9328	-1,9345	-2,0876	-1,3318
8	2,0037	-1,9946	-2,0003	-1,2404
9	1,9998	-1,9998	-1,9998	-1,2498
10	2,0000	-2,0000	-2,0000	-1,2500
11	2,0000	-2,0000	-2,0000	-1,2500

Na primeira solução tem-se: $x^2 - 2x + 2 = 0 \Rightarrow x = 1 \pm \boldsymbol{i}$, na segunda solução tem-se: $x^2 + 2x + 1,25 = 0 \Rightarrow x = -1 \pm \dfrac{\boldsymbol{i}}{2}$.

4.7 Métodos *Quasi*-Newton

Os métodos *quasi*-Newton são essencialmente aproximações da derivada primeira de $f(x)$, necessária no método de Newton-Raphson.

4.7.1 Método da Secante

O método de Newton-secante, ou simplesmente método da secante, baseia-se na aproximação da derivada da função $f(x)$, que aparece no método clássico de Newton-Raphson, pela equação de diferenças finitas à esquerda:

$$f'(x^{(k)}) = \left.\frac{df(x)}{dx}\right|_{x^{(k)}} \approx \frac{\Delta f}{\Delta x} = \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$

Resultando no seguinte processo iterativo:

$$x^{(k+1)} = \frac{f(x^{(k)})x^{(k-1)} - f(x^{(k-1)})x^{(k)}}{f(x^{(k)}) - f(x^{(k-1)})} \text{ para } k = 1, 2, 3, \cdots$$

Sendo, neste caso, necessários dois pontos para iniciar as iterações, $x^{(0)}$ e $x^{(1)}$, pois a equação da reta descrita pelo processo iterativo é secante à curva em dois pontos, ao passo que no método de

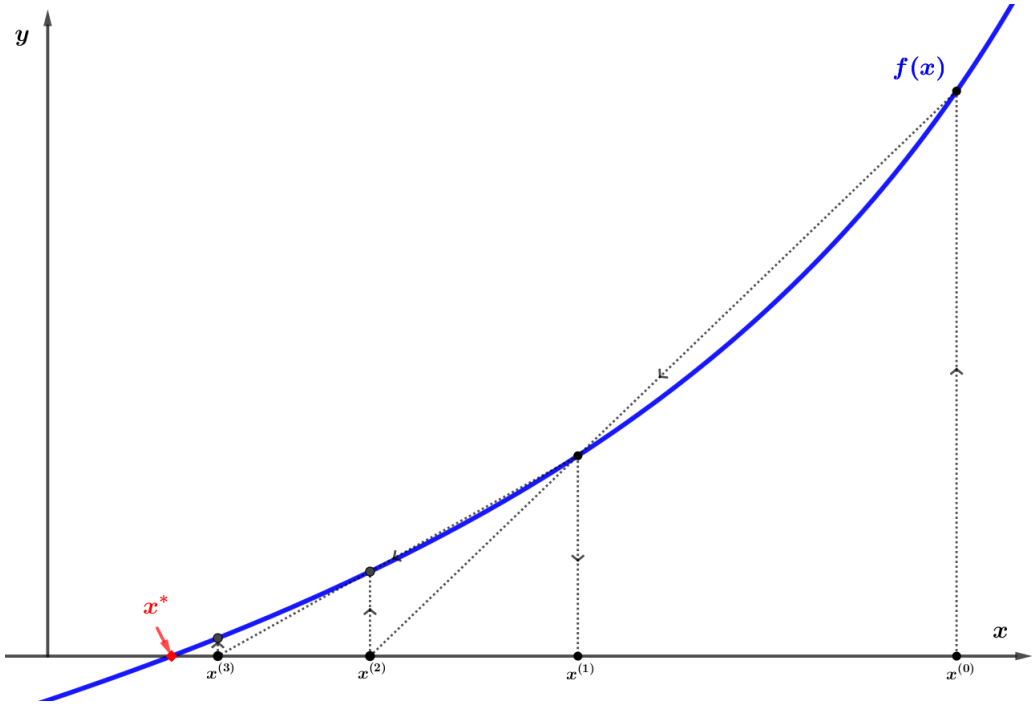


Figura 4.15: Método da secante.

Newton-Raphson convencional a equação da reta é definida por um ponto e a tangente neste ponto. Um esboço gráfico do método é mostrado na Figura 4.15.

A convergência deste método é super-linear, isto é, mais rápida que a convergência linear do método das substituições sucessivas e mais lenta que a convergência quadrática do método de Newton-Raphson, possuindo a seguinte forma:

$$|x^{(k+1)} - x^*| \leq \rho |x^{(k)} - x^*|^{1,618} \text{ sendo } 0 < \rho < 1.$$

A forma algorítmica do método da secante é apresentada a seguir.

Especificação dos valores de ε, δ, k_{maximo}, x^0 e x^1
$k \longleftarrow 0$
$y^0 \longleftarrow f(x^0)$
$y \longleftarrow f(x^1)$
Faça
$$x \longleftarrow \frac{y\,x^0 - y^0\,x^1}{y - y^0}$$
$y^0 \longleftarrow y$
$y \longleftarrow f(x)$
$\Delta \longleftarrow |x - x^1|$
$x^0 \longleftarrow x^1$
$x^1 \longleftarrow x$
$k \longleftarrow k + 1$
enquanto $(\Delta > \varepsilon$ ou $|y| > \delta)$ e $k < k_{maximo}$

Ao final do algoritmo, se $k < k_{maximo}$ então x contém a raiz encontrada de $f(x)$ e y contém o valor de $f(x)$; se o processo parar ao atingir o número máximo de iterações sem convergência, novos valores iniciais x^0 e\ou x^1 devem ser fornecidos.

4.7.2 Método da *Regula-Falsi*

O método da *regula-falsi* ou posição falsa é uma modificação do método da secante, em que a derivada da função $f(x)$ é grosseiramente aproximada pela equação das diferenças finitas em relação a um ponto fixo, $x^{(0)}$:

$$f'(x^{(k)}) = \left.\frac{df(x)}{dx}\right|_{x^{(k)}} \approx \frac{\Delta f}{\Delta x} = \frac{f(x^{(k)}) - f(x^{(0)})}{x^{(k)} - x^{(0)}}$$

Resultando no seguinte processo iterativo:

$$x^{(k+1)} = \frac{f(x^{(k)})\, x^{(0)} - f(x^{0})\, x^{(k)}}{f(x^{(k)}) - f(x^{(0)})} \text{ para } k = 1, 2, 3, \cdots$$

Um esboço gráfico do método é mostrado na Figura 4.16.

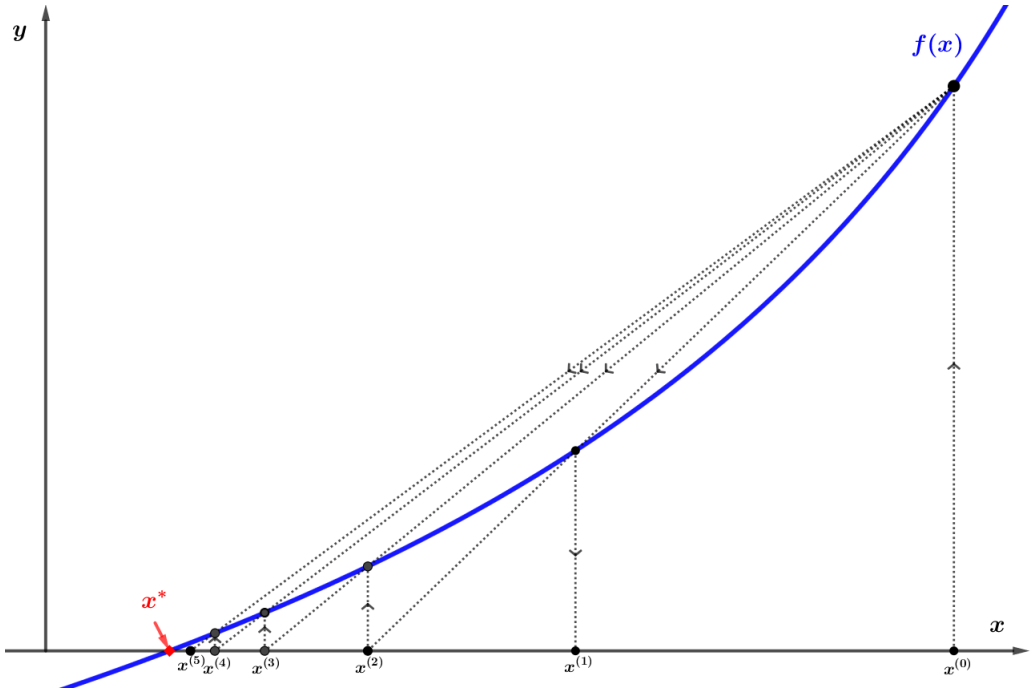


Figura 4.16: Método da *regula-falsi*.

A forma algorítmica do método da *regula-falsi* é apresentada a seguir.

Especificação dos valores de ε, δ, k_{maximo}, x^0 e x^1
$k \longleftarrow 0$
$y^0 \longleftarrow f(x^0)$
$y \longleftarrow f(x^1)$

Faça
$x \longleftarrow \dfrac{y\,x^0 - y^0\,x^1}{y - y^0}$
$y \longleftarrow f(x)$
$\Delta \longleftarrow |x - x^1|$
$x^1 \longleftarrow x$
$k \longleftarrow k + 1$
enquanto $(\Delta > \varepsilon$ ou $|y| > \delta)$ e $k < k_{maximo}$

Ao final do algoritmo, se $k < k_{maximo}$ então x contém a raiz encontrada de $f(x)$ e y contém o valor de $f(x)$; se o processo parar ao atingir o número máximo de iterações sem convergência, novos valores iniciais x^0 e\ou x^1 devem ser fornecidos.

4.7.3 Método de Wegstein

No método de Wegstein[7] ou método da *regula-falsi* modificado, ao invés de manter fixo o ponto base para o cálculo da aproximação da derivada primeira da função $f(x)$, atualiza-se este ponto de acordo com a posição do novo ponto gerado em cada iteração. Assim, o processo iterativo apresenta a seguinte forma:

$$x^{(k+1)} = \frac{f(x_R^{(k)})\, x_L^{(k)} - f(x_L^{(k)})\, x_R^{(k)}}{f(x_R^{(k)}) - f(x_L^{(k)})} \text{ para } k = 0, 1, 2, 3, \cdots$$

Em que, de forma similar ao método da bisseção (Seção 4.2.1):

$$\left[x_L^{(k+1)},\, x_R^{(k+1)}\right] = \begin{cases} \left[x^{(k+1)},\, x_R^{(k)}\right] & \text{se } f(x^{(k+1)})\, f(x_L^{(k)}) > 0 \\ \left[x_L^{(k)},\, x^{(k+1)}\right] & \text{se } f(x^{(k+1)})\, f(x_L^{(k)}) < 0 \end{cases} .$$

Para iniciar o processo iterativo deve-se ter: $f(x_L^{(0)})\, f(x_R^{(0)}) < 0$.

Neste método tem-se $\lambda^{(k)} = \dfrac{f(x_R^{(k)})}{f(x_R^{(k)}) - f(x_L^{(k)})}$, como necessariamente $f(x_L^{(k)})\, f(x_R^{(k)}) < 0$, então $\lambda^{(k)} = \dfrac{f^2(x_R^{(k)})}{f^2(x_R^{(k)}) - f(x_L^{(k)}) f(x_R^{(k)})} = \dfrac{f^2(x_R^{(k)})}{f^2(x_R^{(k)}) + K}$ em que $K > 0$, logo $0 \leq \lambda^{(k)} \leq 1$.

Um esboço gráfico do método é mostrado na Figura 4.17.

Observe que este método é similar ao método da secante quando ocorre alternância de sinal da função $f(x)$ entre iterações sucessivas do método da secante, e similar ao método da *regula-falsi* quando nesse último não ocorre alternância de sinal da função.

O método de Wegstein pode ser implementado pelo mesmo algoritmo descrito na Seção 8.2, para os métodos diretos, em que a função $F(a,b)$ é dada por:

$$F(a,b) = \frac{f_a}{f_a - f_b}.$$

Estes métodos também têm um embasamento geométrico de acordo com o representado na Figura 4.18.

Considerando a semelhança entre os dois triângulos $\widehat{ABC}$ e $\widehat{CDE}$, tem-se:

$$\frac{AB}{BC} = \frac{DE}{CE} \text{ em vista de: } AB = f_a,\, DE = -f_b,\, BC = \lambda(b-a) \text{ e } CE = (1-\lambda)(b-a)$$

o que implica:

$$\frac{f_a}{\lambda(b-a)} = -\frac{f_b}{(1-\lambda)(b-a)} \Rightarrow \lambda = \frac{f_a}{f_a - f_b}.$$

[7] Joseph Henry Wegstein (1922-1985).

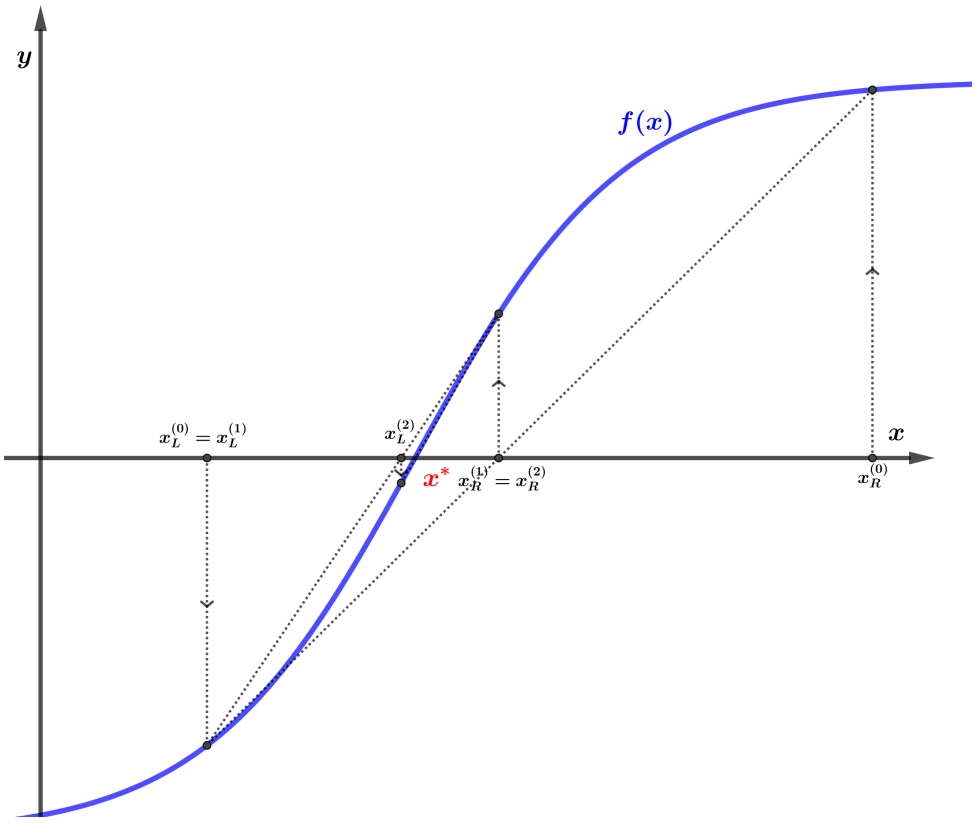


Figura 4.17: Método de Wegstein.

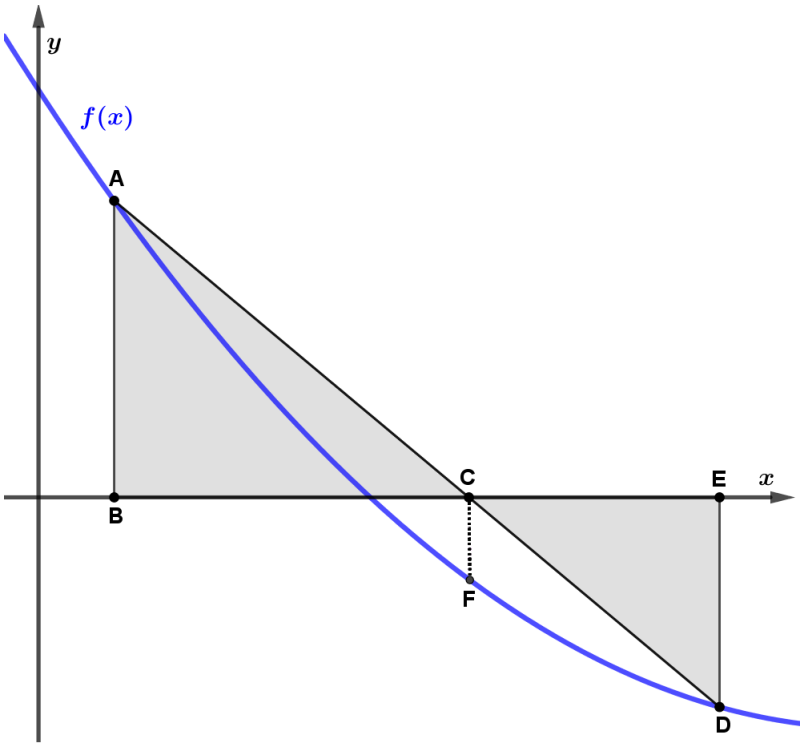


Figura 4.18: Interpretação geométrica do método de Wegstein.

4.8 Método de Müller

O método de Müller[8] pode ser classificado como um método de busca direta (sem cômputo de derivadas), eficiente para a determinação de raízes reais múltiplas e raízes complexas de funções, especialmente polinômios. Este método se fundamenta na busca de raízes de interpolações parabólicas sucessivas da função que se deseja determinar as raízes, assim o processo se inicia após a especificação de três valores da variável independente x, ordenados de forma decrescente do valor absoluto da função $f(x)$ avaliada nesses pontos, ou seja, x_2 é o ponto mais próximo à raiz:

$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}$. Calculando-se a seguir: $\mathbf{y} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}$, com esses três pontos calculam-se:

ETAPA A

$$\mathbf{h} = \begin{pmatrix} x_1 - x_0 \\ x_2 - x_1 \end{pmatrix},\ \Delta = \begin{pmatrix} \Delta_0 \\ \Delta_1 \end{pmatrix} = \begin{pmatrix} \dfrac{y_1 - y_0}{h_0} \\ \dfrac{y_2 - y_1}{h_1} \end{pmatrix} \text{ e } f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{h_0 + h_1}.$$

Com estes valores é possível calcular o polinômio interpolador quadrático de acordo com:

$$p_2(x) = f(x_2) + f[x_1, x_2](x - x_2) + f[x_0, x_1, x_2](x - x_1)(x - x_2),$$

como esquematizado na Figura 4.19.

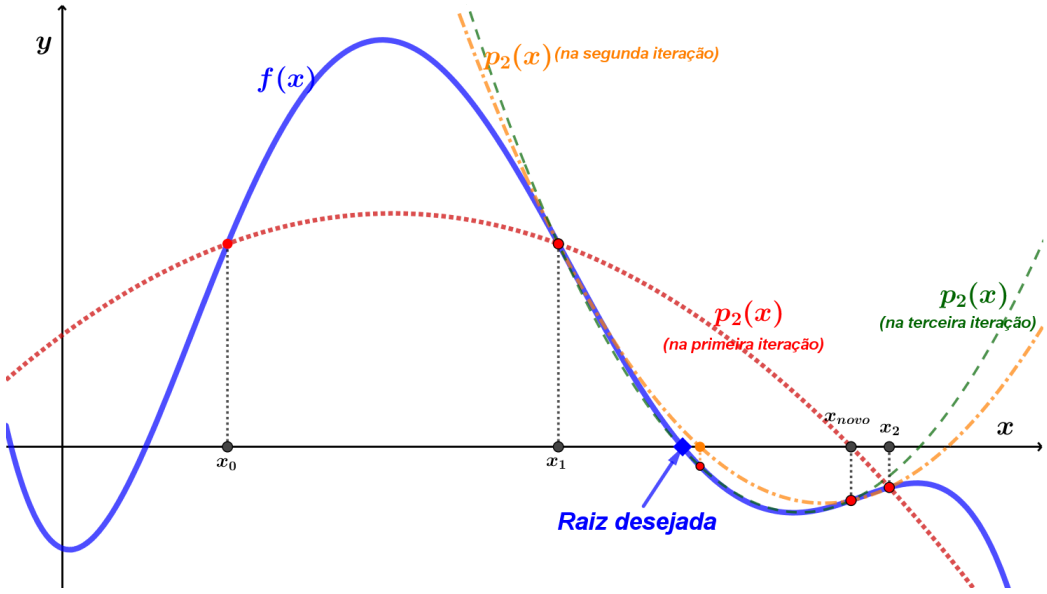


Figura 4.19: Método de Müller.

Reescrevendo o polinômio interpolador em termos da variável $z = x - x_2$, resulta:

$$p_2(z) = f(x_2) + f[x_1, x_2]z + f[x_0, x_1, x_2](z + h_1)z = f[x_0, x_1, x_2]z^2 + \{f[x_0, x_1, x_2] + f[x_1, x_2]\}z + f(x_2).$$

Identificando $a = f[x_0, x_1, x_2]$, $b = f[x_0, x_1, x_2] + f[x_1, x_2]$ e $c = f(x_2)$, as raízes de $p_2(z)$ são dadas por:

$$\begin{cases} r_1 = -\left(\dfrac{b + \sqrt{b^2 - 4ac}}{2a}\right) \\ r_2 = -\left(\dfrac{b - \sqrt{b^2 - 4ac}}{2a}\right) \end{cases}.$$

[8] David Eugene Müller (1924–2008).

Como em trechos *lineares* de $f(x)$ os valores de $a = f[x_0, x_1, x_2]$ aproximam-se de zero, a maneira mais conveniente de expressar as raízes de $p_2(z)$, baseado no fato que $r_1\, r_2 = \dfrac{c}{a}$, é na forma:

$$\begin{cases} r_1 = \dfrac{c}{a r_2} = -\left(\dfrac{2c}{b - \sqrt{b^2 - 4ac}}\right) \\ r_2 = \dfrac{c}{a r_1} = -\left(\dfrac{2c}{b + \sqrt{b^2 - 4ac}}\right) \end{cases} .$$

Como essas raízes estão expressas em termos da variável $z = x - x_2$, opta-se para calcular o próximo valor da variável independente x por:

$$x_{novo} = x_2 - \left(\frac{2c}{b \pm \sqrt{b^2 - 4ac}}\right),$$

com o sinal '+' ou '−' escolhido de modo que: $|b \pm \sqrt{b^2 - 4ac}|$ assuma o maior valor possível, fazendo assim com que x_{novo} esteja mais *próximo* de x_2.

A seguir novos valores dos pontos interpoladores são escolhidos de acordo com:

$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_{novo} \end{pmatrix}$ e $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ f(x_{novo}) \end{pmatrix}$, o que equivale a *descartar* o ponto x_0 e reordenar os componentes dos vetores $\mathbf{x}$ e $\mathbf{y}$, de forma decrescente do valor absoluto da função $f(x)$. A seguir, o procedimento é reiniciado pela **ETAPA A** e assim repetido até $|x_{novo} - x_2| < \varepsilon$ e $|f(x_{novo})| < \delta$, sendo ε e δ especificados pelo usuário.

Na tabela a seguir, estão listados os valores obtidos pela aplicação do método de Müller na determinação da raiz de $f(x) = e^{-x^2} - x$, cuja solução é $x^* = 0,652919$.

Iteração	**0**	**1**	**2**	**3**	**4**
x_0	0,000000	0,500000	1,000000	0,664528	0,652728
x_1	0,500000	1,000000	0,664528	0,652728	0,652918
x_2	1,000000	0,664528	0,652728	0,652918	0,652919
y_0	1,000000	0,278801	-0,632121	-0,021519	0,000352
y_1	0,278801	-0,632121	-0,021519	0,000352	0,0000004
y_2	-0,632121	-0,021519	0,000352	0,0000004	0,000000

4.9 Critérios de Convergência

A avaliação da convergência do processo iterativo é feita através do cômputo do erro da variável x na iteração k, devendo ser menor que um valor especificado pelo usuário, em acordo com $\varepsilon^{(k)} = |x^{(k)} - x^*| \leq \varepsilon$. No entanto, o valor da solução do problema x^* não é conhecido e a avaliação da convergência é feita segundo diferentes critérios, que são apresentados e discutidos a seguir.

- **Critério do erro absoluto em** $x^{(k+1)}$:
  $|x^{(k+1)} - x^{(k)}| \leq \varepsilon_{abs}$. Desvantagem: requer um conhecimento da ordem de grandeza de x^*.
  Uma melhoria na avaliação deste critério pode ser feita buscando relacioná-lo com o erro real $e^{(k)} = |x^{(k)} - x^*|$. Para isto, introduz-se o conceito de **taxa de convergência** expresso por: $|x^{(k+1)} - x^{(k)}| \leq q|x^{(k)} - x^{(k-1)}|$, em que q é suposto constante ao longo de todo processo iterativo e $0 < q < 1$. Assim:
  $|x^{(k+2)} - x^{(k+1)}| \leq q|x^{(k+1)} - x^{(k)}| \leq q^2|x^{(k)} - x^{(k-1)}|$, ou seja:
  $|x^{(k+j)} - x^{(k+j-1)}| \leq q^j|x^{(k)} - x^{(k-1)}|$.
  Usando a desigualdade triangular: $|a - b| \leq |a - c| + |c - b|$, tem-se:

$$|x^{(k+m)} - x^{(k)}| \leq \sum_{j=1}^{m} |x^{(k+j)} - x^{(k+j-1)}| = \left(\sum_{j=1}^{m} q^j \right) |x^{(k)} - x^{(k-1)}|.$$

Se o processo iterativo convergir para x^* tem-se: $\lim_{m \longrightarrow \infty} x^{(k+m)} = x^*$ e $\sum_{j=1}^{\infty} q^j = \frac{q}{1-q}$ (série geométrica), permitindo relacionar:
$|x^* - x^{(k)}| \leq \frac{q}{1-q} |x^{(k)} - x^{(k-1)}|$.
Assim, impondo $\frac{q}{1-q} |x^{(k)} - x^{(k-1)}| \leq \varepsilon$, tem-se: $|x^* - x^{(k)}| \leq \varepsilon$.
A maior dificuldade deste procedimento é a estimação do parâmetro q, seu valor deve ser *atualizado* em cada iteração k. Uma boa estimativa para esse valor pode ser obtida a partir da expressão:

$$|x^{(k+1)} - x^{(k)}| \leq q^k |x^{(1)} - x^{(0)}| \Rightarrow q \approx \left| \frac{x^{(k+1)} - x^{(k)}}{x^{(1)} - x^{(0)}} \right|^{\left(\frac{1}{k}\right)}.$$

Sugerindo-se considerar na primeira iteração a estimativa conservadora $q = 0,99$, o que equivale a considerar: $\frac{0,99}{0,01} |x^{(1)} - x^{(0)}| = 99 |x^{(1)} - x^{(0)}| \leq 100 |x^{(1)} - x^{(0)}| \leq \varepsilon$, ou seja, $|x^{(1)} - x^{(0)}| \leq \frac{\varepsilon}{100}$.

- **Critério do erro relativo em** $x^{(k+1)}$:
  $|x^{(k+1)} - x^{(k)}| \leq \varepsilon_{rel} |x^{(k)}|$. Desvantagem: falhar quando $x^* \approx 0$.

- **Critério do erro absoluto em** $f(x^{(k+1)})$:
  $|f(x^{(k+1)})| \leq \delta$. Desvantagem: pode mascarar a convergência na variável x, ilustrado na Figura 4.20.

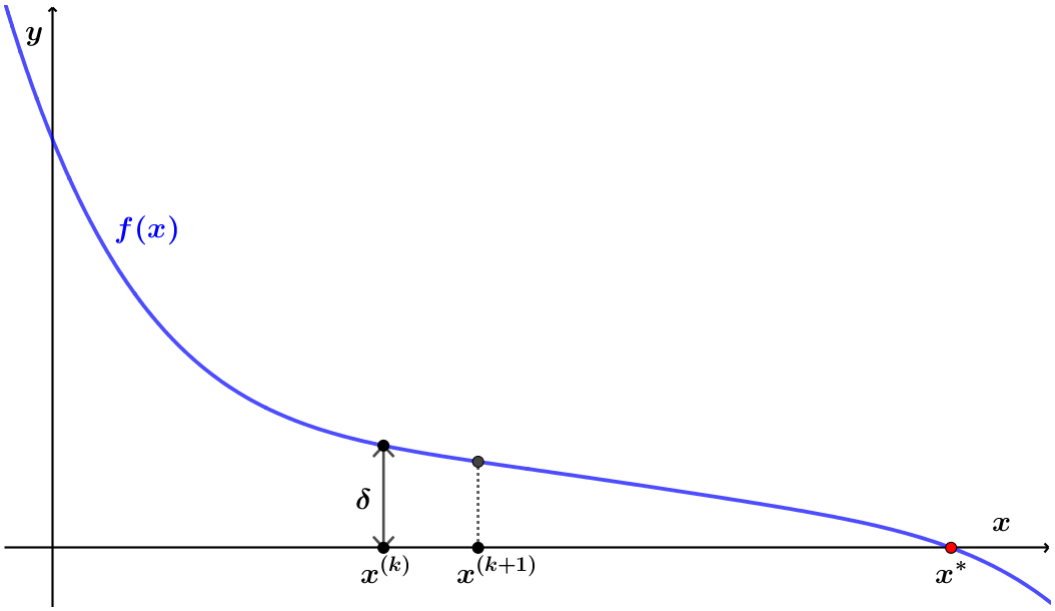


Figura 4.20: Critério do erro absoluto em $f(x^{(k+1)})$.

- **Combinação dos critérios do erro relativo e absoluto em** $x^{(k+1)}$: $|x^{(k+1)} - x^{(k)}| \leq \varepsilon_{rel} |x^{(k)}| + \varepsilon_{abs}$. Desvantagem: persiste a necessidade de conhecer a ordem de grandeza de x^*.

Ordem de Convergência e Coeficiente Assintótico de Convergência. Seja um sequência

$x^{(0)}, x^{(1)}, x^{(2)}, \cdots, x^{(n)}$ que converge para x^* e seja $\varepsilon^{(n)} = x^* - x^{(n)}$. Se existe um número α e uma constante ρ tais que $\lim_{n\longrightarrow\infty} \frac{|\varepsilon^{(n+1)}|}{|\varepsilon^{(n)}|^\alpha} = \rho$, então α é chamada de ordem de convergência da sequência e ρ é a coeficiente assintótico de convergência.

Como exemplo desses conceitos tem-se:

1. Método das substituições sucessivas: $\rho = |g'(x^*)|$ e $\alpha = 1$ *Método de Convergência Linear*;
2. Método de Newton-Raphson: $\rho = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$ e $\alpha = 2$ *Método de Convergência Quadrática*;
3. Método da Secante: Tanto no método de Newton-Raphson como no método da secante, aproxima-se em cada iteração k a função $f(x)$ por uma reta, assim:
   $p_1(x) = f(a) + f[a,b](x-a)$ sendo no método de Newton-Raphson $a = b = x^{(k)}$ e no método da secante $a = x^{(k)}$ e $b = x^{(k-1)}$. Lembrando que
   $f[a,b] = \frac{f(b)-f(a)}{b-a}$ e $f[a,a] = \lim_{b\longrightarrow a}\left(\frac{f(b)-f(a)}{b-a}\right) = f'(a)$.
   Em ambos os métodos calcula-se o próximo ponto $\hat{x}$ tal que
   $p_1(\hat{x}) = 0 = f(a) + f[a,b](\hat{x}-a) \Rightarrow \hat{x} = a - \frac{f(a)}{f[a,b]}$.
   No método das diferenças divididas de Newton (Seção 3.3) viu-se que:
   $f(x) = f(a) + f[a,b](x-a) + f[a,b,x](x-a)(x-b)$, equação que é válida para qualquer valor de x, em particular se $x = x^*$ (solução de $f(x^*) = 0$), tem-se:
   $f(x^*) = 0 = f(a) + f[a,b](x^*-a) + f[a,b,x^*](x^*-a)(x^*-b)$ ou seja:

$$x^* = \left(a - \frac{f(a)}{f[a,b]}\right) - \frac{f[a,b,x^*]}{f[a,b]}(x^*-a)(x^*-b) = \hat{x} - \frac{f[a,b,x^*]}{f[a,b]}(x^*-a)(x^*-b).$$

   No método de Newton-Raphson
   $a = b = x^{(k)} \Rightarrow f[a,a] = f'(a) = f'(x^{(k)})$ e $f[a,a,x^*] = f[x^{(k)},x^{(k)},x^*] = \frac{f''(\eta^{(k)})}{2}$ em que $\eta^{(k)}$ é algum valor de x entre $x^{(k)}$ e x^*.
   Identificando: $\hat{x} = x^{(k+1)}$, $x^* - x^{(k+1)} = \varepsilon^{(k+1)}$, $(x^*-a)(x^*-b) = (x^*-x^{(k)})^2 = (\varepsilon^{(k)})^2$ chega-se a:

$$\varepsilon^{(k+1)} = -\frac{f''(\eta^{(k)})}{2f'(x^{(k)})}(\varepsilon^{(k)})^2$$

   Reconfirmando que o método de Newton-Raphson é um método de convergência quadrática.
   No método da secante:
   $a = x^{(k)}$ e $b = x^{(k-1)}$, assim: $x^* - a = x^* - x^{(k)} = \varepsilon^{(k)}$, $x^* - b = x^* - x^{(k-1)} = \varepsilon^{(k-1)}$, $f[a,b] = f[x^{(k)},x^{(k-1)}] = f'(\eta^{(k)})$ e $f[x^{(k)},x^{(k-1)},x^*] = \frac{1}{2}f''(\sigma^{(k)})$. Sendo $\eta(k)$ e $\sigma(k)$ valores de x entre $x^{(k)}, x^{(k-1)}$ e x^*. Resultando em:

$$\varepsilon^{(k+1)} = -\frac{f''(\sigma^{(k)})}{2f'(\eta^{(k)})}\varepsilon^{(k)}\varepsilon^{(k-1)}$$

   Quando $k \longrightarrow \infty$ tem-se $\eta^{(k)} \longrightarrow x^*$ e $\sigma^{(k)} \longrightarrow x^*$ assim:

$$\varepsilon^{(k+1)} \approx -\frac{f''(x^*)}{2f'(x^*)}\varepsilon^{(k)}\varepsilon^{(k-1)}.$$

   Logo:
   $|\varepsilon^{(k+1)}| = \rho^{(k)}|\varepsilon^{(k)}\varepsilon^{(k-1)}|$ com $\rho^{(k)} = \left|\frac{f''(\sigma^{(k)})}{2f'(\eta^{(k)})}\right|$ e $\lim_{k\longrightarrow\infty}\rho^{(k)} = \rho_\infty = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$. Para encontrar a ordem de convergência e o coeficiente assintótico de convergência do método da secante, procede-se da seguinte forma:

$$\frac{|\varepsilon^{(n+1)}|}{|\varepsilon^{(n)}|^{\alpha}} = \rho^{(n)}|\varepsilon^{(n)}|^{1-\alpha}|\varepsilon^{(n-1)}| = \rho^{(n)}\left|\frac{|\varepsilon^{(n)}|}{|\varepsilon^{(n-1)}|^{\alpha}}\right|^{p} \Rightarrow p = 1-\alpha \text{ e } 1 = -\alpha\cdot p.$$

Assim: $1 = \alpha(\alpha-1)$, $\alpha^2-\alpha-1=0$ cuja raiz positiva é $\alpha = \dfrac{1+\sqrt{5}}{2} = 1{,}618034$ e $p = -\dfrac{1}{\alpha} = 1-\alpha = \dfrac{1-\sqrt{5}}{2} = -0{,}618034$.

Definindo: $y^{(n)} = \dfrac{|\varepsilon^{(n)}|}{|\varepsilon^{(n-1)}|^{\alpha}}$ e $y^{(n+1)} = \dfrac{|\varepsilon^{(n+1)}|}{|\varepsilon^{(n)}|^{\alpha}} \Rightarrow y^{(n+1)} = \rho^{(n)}(y^{(n)})^{p}$.

Na convergência: $y^{(n)} = y^{(n+1)} = y_{\infty}$ e $\rho^{(n)} = \rho_{\infty} = \left|\dfrac{f''(x^*)}{2f'(x^*)}\right|$, sendo:

$y_{\infty} = \rho_{\infty}(y_{\infty})^{p} \Rightarrow y_{\infty} = (\rho_{\infty})^{\frac{1}{\alpha}}$.

Mostrando que para n elevado $y^{(n)} = \dfrac{|\varepsilon^{(n)}|}{|\varepsilon^{(n-1)}|^{\alpha}} \approx \left|\dfrac{f''(x^*)}{2f'(x^*)}\right|^{\frac{1}{\alpha}}$ em que $\alpha = \dfrac{1+\sqrt{5}}{2} = 1{,}618034$ e $\dfrac{1}{\alpha} = \alpha - 1 = \dfrac{\sqrt{5}-1}{2} = 0{,}618034$. O que permite afirmar que o método da secante apresenta ordem de convergência igual a 1,618034 e coeficiente assintótico de convergência igual a $\left|\dfrac{f''(x^*)}{2f'(x^*)}\right|^{\frac{1}{\alpha}}$. O que demonstra que o método da secante converge mais rapidamente que o método das substituições sucessivas, porém mais lentamente que o método de Newton-Raphson.

4.10 Problemas Propostos

Problema 4.1 Aplique o método de Newton-Raphson na determinação da raiz real positiva da função com 6 algorismo significativos corretos:

$f(x) = e^{-x} - 2\sqrt{x}$ no intervalo $[0,1]$. Explique porque não é conveniente utilizar como condição inicial $x(0) = 0$.

Compare a velocidade de convergência do método com a obtida pelo método da bisseção. Refaça o problema pela aplicação de uma forma asseguradamente convergente do método das substituições sucessivas.

Problema 4.2 Uma modificação do método de Weigstein é proposta em que se utiliza uma nova parametrização do intervalo de busca em acordo com a expressão:

$$x = \frac{a+b}{2} + \lambda\frac{b-a}{2} \text{ em que } -1 \le \lambda \le +1.$$

Nessa nova forma, verifica-se:

(i) $\lambda = -1 \Rightarrow x = a$: extremidade inferior do intervalo de busca;

(ii) $\lambda = 0 \Rightarrow x = \dfrac{a+b}{2}$: ponto médio do intervalo de busca;

(iii) $\lambda = +1 \Rightarrow x = b$: extremidade superior do intervalo de busca.

Para determinar o valor de λ em cada iteração se utiliza uma interpolação quadrática inversa fundamentada nos três pontos:

x	a	$\dfrac{a+b}{2}$	b
y	y_a	y_m	y_b
λ	-1	0	+1

A interpolação quadrática inversa é descrita por:

$$\lambda(y) = p_2(y) = \frac{(y-y_b)(y-y_m)}{(y_b-y_a)(y_a-y_m)} + \frac{(y-y_a)(y-y_m)}{(y_b-y_a)(y_b-y_m)}.$$

Calculando o valor de λ em cada iteração por:

$$\lambda = \lambda(0) = p_2(0) = \frac{y_b\, y_m}{(y_b - y_a)(y_a - y_m)} + \frac{y_a\, y_m}{(y_b - y_a)(y_b - y_m)}.$$

A forma algorítmica de implementação do método é descrita a seguir.

$f_a \longleftarrow f(a)$
$f_b \longleftarrow f(b)$
Se $f_a\, f_b > 0$ então busque novos valores de a e b
$k \longleftarrow 0$

Faça

$$y_m \longleftarrow f\left(\frac{a+b}{2}\right)$$
$$\lambda \longleftarrow \frac{y_b\, y_m}{(y_b - y_a)(y_a - y_m)} + \frac{y_a\, y_m}{(y_b - y_a)(y_b - y_m)}$$
$$x \longleftarrow \frac{a+b}{2} + \lambda\frac{b-a}{2}$$
$$y \longleftarrow f(x)$$
$$\text{Se } y\, f_a > 0 \text{ faça } \begin{cases} f_a \longleftarrow y \\ a \longleftarrow x \end{cases}$$
$$\text{senão: } \begin{cases} f_b \longleftarrow y \\ b \longleftarrow x \end{cases}$$
$$\Delta \longleftarrow |b - a|$$
$$k \longleftarrow k + 1$$

enquanto $(\Delta > \varepsilon \text{ ou } |y| > \delta)$ e $k < k_{max}$

(a) Existe alguma limitação à aplicação deste novo procedimento? Qual? Justifique sua resposta.
(b) Aplique o procedimento na determinação da raiz real positiva da função: $f(x) = e^{-x} - 2x^2$ no intervalo $[0, 1]$. Baseado nos resultados obtidos, compare o desempenho do novo método com o do método de Weigstein convencional.

Problema 4.3 Em problemas de transferência de calor é importante calcular as raízes reais positivas da equação: $\dfrac{\text{tg}(\lambda)}{\lambda} = K$ em que K é uma constante real conhecida, note que (exceto no caso em que $K = 1$) como $\lim\limits_{\lambda \longrightarrow 0}\left(\dfrac{\text{tg}(\lambda)}{\lambda}\right) = 1$, $\lambda = 0$ não é raiz da equação.

Para evitar as descontinuidades da função tangente reescreve-se a equação original na forma: $f(\lambda) = \dfrac{\text{sen}(\lambda)}{\lambda} - K\cos(\lambda) = 0$, note que: $f(0) = 1 - K \neq 0$ se $K \neq 1$.

Apresente o procedimento iterativo que traduza o método de Newton-Raphson aplicado à equação e a seguir determine a primeira raiz positiva da equação, com uma acurácia até a sexta casa decimal, quando $K = 2$, mostrando claramente em seu procedimento como é evitada a possível descontinuidade em $\lambda = 0$. Sugira e implemente uma metodologia para determinar as 10 primeiras raízes reais positivas da equação.

Problema 4.4 Uma reação química de primeira ordem, irreversível e em fase líquida é conduzida em um reator tanque operando de forma cíclica. Assim, conduz-se uma batelada por um tempo t e, depois de transcorrido esse tempo, esvazia-se e se limpa o reator, levando essa última operação um tempo igual a t_c. Após passar pela etapa de esvaziamento/limpeza conduz-se novamente uma batelada por um tempo t, e assim sucessivamente. Pode-se demonstrar que o tempo ótimo da fase batelada que maximiza a taxa de produção do processo é obtido através da equação não linear:

$$[1 + k(t + t_c)]e^{-k\,t} = 1 \quad (1)$$

ou na forma logarítmica:

$$k\,t = \ln[1 + k(t + t_c)] \quad (2)$$

Sendo: $\begin{cases} k \text{ constante de velocidade da reação} = 2{,}5h^{-1}; \\ t_c \text{ tempo da operação de esvaziamento/limpeza} = 0{,}5h. \end{cases}$

Para calcular t dois procedimentos iterativos são propostos:

(a) De (1) obtém-se: $t = \dfrac{e^{k\,t} - 1}{k} - t_c$ sugerindo o procedimento recursivo:

$$t^{(j+1)} = \frac{e^{k\,t^{(j)}} - 1}{k} - t_c \text{ para } j = 0, 1, \cdots \text{ com } t^{(0)} = t_c;$$

(b) De (2) obtém-se: $t = \dfrac{\ln[1 + k(t + t_c)]}{k}$ sugerindo o procedimento recursivo:

$$t^{(j+1)} = \frac{\ln[1 + k(t^{(j)} + t_c)]}{k} \text{ para } j = 0, 1, \cdots \text{ com } t^{(0)} = t_c;$$

O procedimento (a) não converge, enquanto que o procedimento (b) converge à solução do problema (para o conjunto de parâmetros utilizados) $t^* = 0{,}50155\ h$ após 6 iterações. Explique porque isto ocorre. Refaça o problema aplicando o método de Newton-Raphson às duas formulações, discuta e confronte seus resultados entre si e com os dois métodos de substituições sucessivas anteriores.

Problema 4.5 Sugere-se o seguinte procedimento para determinar iterativamente as raízes de uma função não linear em uma variável $f(x) = 0$: buscam-se os mínimos locais de $g(x) = [f(x)]^2$, isto é os valores de x que anulam a derivada de $g(x)$, ou seja, buscam-se as raízes de uma nova função:

$$F(x) = g'(x) = 2f(x)f'(x).$$

Aplique o método de Newton-Raphson à função $F(x)$ e mostre o algoritmo recursivo correspondente e comparando-o com o obtido aplicando diretamente o método de Newton-Raphson à função original.

Ilustre o processo iterativo, em ambos os casos, aplicando-o a uma função de sua escolha, por exemplo: $f(x) = x^2 - 2$ ou $f(x) = e^{-x} - x$. Analise e comente os resultados obtidos, apresentando as vantagens e/ou desvantagens do procedimento sugerido.

Problema 4.6 Determine todas as raízes dos seguintes polinômios com seis algarismos significativos corretos:

$$\begin{aligned} P_4(x) &= x^4 + 3x^3 - 3x^2 + 32x + 180 \\ P_6(x) &= x^6 - 2x^5 + 2x^4 + x^3 + 6x^2 - 6x + 8 \\ P_7(x) &= x^7 + 4x^6 + 10x^5 - 8x^4 - 21x^3 + 56x^2 + 10x - 52 \\ P_{10}(x) &= 184756x^{10} - 923780x^9 + 1969110x^8 - 2333760x^7 + 1681680x^6 + \\ &- 756756x^5 + 210210x^4 - 34320x^3 + 2970x^2 - 110x + 1 \end{aligned}$$

Problema 4.7 Obtenha a solução dos problemas abaixo com seis algarismos significativos corretos.

(a) Resolva $x = \cos(x)$ por substituições sucessivas considerando $x^{(0)} = 1$.

(b) Mostre que $x = \cos(x)$ pode ser transformado em $x = 1 - \dfrac{[\text{sen}(x)]^2}{1 + x}$, verificando em quantas iterações obtém-se a mesma solução do problema anterior com a mesma condição inicial.

(c) Mostre que $x = \cos(x)$ pode ser transformado em $x = \sqrt{x\cos(x)}$, verificando em quantas iterações obtém-se a mesma solução dos problemas anteriores com a mesma condição inicial.

(d) Esboce o gráfico da função $f(x) = \text{sen}(x) - \text{cotg}(x) = 0$ e resolva a equação pelo método de Newton-Raphson considerando $x^{(0)} = 1$ (note que o problema apresenta solução analítica).

(e) Resolva o problema anterior pelo método da secante adotando $x^{(0)} = 0{,}5$ e $x^{(1)} = 1$. Compare os novos resultados com os do problema anterior.

(f) Desenvolva um procedimento iterativo baseado no método de Newton-Raphson que permita calcular as raízes cúbicas de números reais, ilustre seu procedimento no cálculo de $\sqrt[3]{7}$, considerando $x^{(0)} = 2$. Aplique o procedimento desenvolvido na determinação do par de raízes complexas solução problema. (Note que no problema $p_3(x) = x^3 - \alpha = 0$ se $\alpha > 0$ há apenas uma troca de sinais nos coeficientes e nenhuma troca de sinal em $p_3(-x)$ o que indica, pela regra de sinais de Descartes, que há apenas uma raiz real positiva, aplicando o mesmo raciocínio para $\alpha < 0$ verifica-se que neste caso há apenas uma raiz real negativa. Como o polinômio é de terceiro grau e de coeficientes reais as duas raízes restantes são um par conjugado de raízes complexas).

(g) Resolva $e^x + x^4 + x = 2$ pelo método da bisseção no intervalo [0, 1].

(h) Encontre a solução real da equação $x^3 = 5x + 6$ pelos métodos da *regula falsi* e de Weigstein, adotando $a = 2$ e $b = 5$.

Problema 4.8 Para o escoamento turbulento de um fluido em um tubo liso, a expressão abaixo pode ser usada para determinar o fator de atrito, f, em função do número de Reynolds, Re.

$$\sqrt{f} = f\left[1,74\ln(Re\sqrt{f}) - 0,4\right]$$

(a) Calcule o fator de atrito para $Re = 5000$, usando o método de Newton-Raphson, com seis algarismos significativos corretos. Uma boa estimativa inicial para o fator de atrito pode ser obtido pela equação de Blasius: $f = 0,316Re^{-0,25}$;

(b) Verifique se a solução pode ser obtida aplicando diretamente o método das substituições sucessivas isolando o fator de atrito, f, no lado direito da equação, justifique a adequação ou não do procedimento.

Problema 4.9 Uma bomba centrífuga é empregada no escoamento de água de um ponto à pressão atmosférica a outro ponto também à pressão atmosférica, porém a uma altura h. Para calcular a vazão volumétrica de água deve se resolver o seguinte sistema de equações não lineares:

(a) Equação da bomba centrífuga: $p_2 - p_{atm} = a - bQ^{\frac{3}{2}}$;

(b) Perda de carga ao longo da tubulação e elevação da água:

$$p_2 - p_{atm} + \rho gh = 8\frac{f\rho LQ^2}{\pi^2 D^5}.$$

Os seguintes valores da propriedades e dos parâmetros das equações são conhecidos: $f = 0,03$ (adimensional), $a = 11\,atm$, $b = 1,5\dfrac{atm\,s^{1,5}}{m^{4,5}}$, $h = 30\,m$, $\rho = 997,1\dfrac{kg}{m^3}$, $L = 500\,m$ e $D = 5,08\,cm$ (2 polegadas).

Calcule a vazão Q e a pressão p_2. (Considere $g = 9,8\dfrac{m}{s^2}$ e $1\,atm = 101325\,Pa$).

Problema 4.10 Para o cálculo da viscosidade do orto-xileno propõe-se o emprego da seguinte expressão:

$$\ln[\mu(T)] = -3,332 + \frac{1,039\times 10^3}{T} - 1,768\times 10^{-3}T + 1,076\times 10^{-6}T^2$$

em que T é a temperatura em Kelvin e μ é a viscosidade em centipoise, tal expressão é válida na faixa $245 \leq T \leq 620\,K$.

Calcule a temperatura na qual a viscosidade do orto-xileno é a metade da viscosidade a $400\,K$. Indique explicitamente o procedimento adotado e os valores da temperatura em todas as iterações do processo recursivo. (Utilize o valor de $600\,K$ como chute inicial e considere uma tolerância relativa de 10^{-4}).

Problema 4.11 O Modelo de Wagner para o cálculo da pressão de vapor de substâncias puras é dado pela expressão:

$$\ln(P_v/P_c) = \frac{a\,x + b\,x^{1,5} + c\,x^3 + d\,x^6}{1-x}$$

em que:
$$\begin{cases} x = 1 - \dfrac{T}{T_c} \\ T: \text{ temperatura em } K \\ T_c: \text{ temperatura crítica em } K \\ P_c: \text{ pressão crítica em } bar \\ P_v: \text{ pressão de vapor em } bar \end{cases}$$

Quando $P_v = P$ (pressão ambiente), T é a temperatura de saturação da substância. Calcule a temperatura de saturação da água para $P = 1\,bar$ com quatro algarismos significativos corretos, dados: $a = -7,76$; $b = 1,46$; $c = -2,78$; $d = -1,23$; $P_c = 221,2\,bar$ e $T_c = 647,3\,K$.

Indique claramente em seus resultados o valor inicial considerado e os valores da temperatura em todas as iterações.

Problema 4.12 A equação de estado de Van der Waals é descrita por:

$$\left(P + \frac{a}{v^2}\right)(v - b) = RT$$

em que:
$$\begin{cases} P: \text{ pressão em } atm \\ v: \text{ volume específico molar em } \dfrac{litro}{mol} \\ R: \text{constante universal dos gases } 0,082054\ \dfrac{litro\ atm}{mol\ K} \\ a: \text{ constante específica do gás } \dfrac{(litro)^2 atm}{(mol)^2} \\ b: \text{ constante específica do gás } \dfrac{litro}{mol} \end{cases}$$

Os valores das constantes a e b para diferentes gases são tabelados a seguir:

Gás	$a\left(\dfrac{(litro)^2\ atm}{(mol)^2}\right)$	$b\left(\dfrac{litro}{mol}\right)$
Gás carbônico	3,592	0,04267
Anilina dimetílica	37,490	0,19700
Hélio	0,03412	0,02370
Óxido nítrico	1,340	0,02789

Sabendo-se que a temperatura crítica do gás (temperatura acima da qual o gás não pode se liquefazer) é dada por: $T_c = \dfrac{8a}{27Rb}$ resolva o problema, para cada um dos gases tabelados, adotando $T > T_c$ e neste caso mostre que para qualquer pressão há apenas uma solução da equação. Adotando $T < T_c$ adote valores de P em que há apenas uma solução e valores de P em que a equação apresenta três soluções, neste último caso: $v_1 < v_2 < v_3$ sendo v_1 o volume específico molar da fase líquida, v_3 o volume específico molar da fase gás e v_2 solução que não apresenta significado físico. Mostre também como calcular a faixa de pressão dentro da qual o sistema apresenta três soluções.

Sugestão: reescreva a equação de Van der Waals em termos das seguintes variáveis adimensionais:
$$\begin{cases} \theta = \dfrac{T}{T_c} \text{ (temperatura adimensional)} \\ \upsilon = \dfrac{V}{b} \text{ (volume específico molar adimensional)} \\ p = \dfrac{P\,b^2}{a} \text{ (pressão adimensional)} \end{cases}.$$

Problema 4.13 F moles por hora de gás natural são alimentados continuamente em um vaso de *flash*, como representado na Figura 4.21.

A operação estacionária deste vaso é descrita pelos balanços:

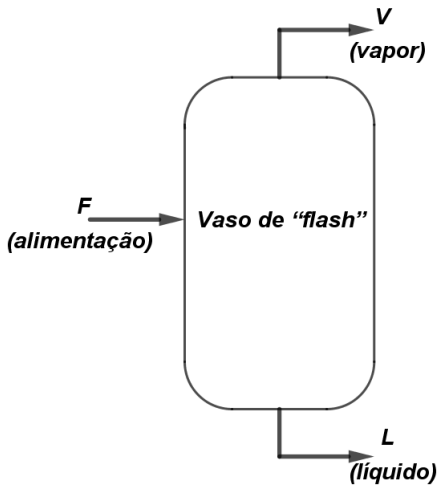


Figura 4.21: Diagrama de um vaso de *flash*.

- Balanço global: $F = L + V$
- Balanço por componente: $Fz_i = Lx_i + Vy_i$ para $i = 1, 2, \cdots, n$
- Relação de equilíbrio (suposta linear): $y_i = K_i\, x_i$ para $i = 1, 2, \cdots, n$
- Restrições algébricas (decorrentes das definições de frações molares):

$$\sum_{i=1}^{n} x_i = 1 \text{ e } \sum_{i=1}^{n} y_i = 1.$$

Mostre que com a manipulação dessas equações obtém-se a equação não linear de Rachford-Rice:

$$f(\phi) = \sum_{i=1}^{n} \left(\frac{K_i z_i}{\phi(K_i - 1) + 1} \right) - 1 = 0, \text{ em que } \phi = V/F.$$

Sabendo-se que as composições de alimentação e constantes de equilíbrio, à temperatura do vaso, são conhecidas e tabeladas abaixo, determine os valores de ϕ, x_i e y_i para $i = 1, 2, \cdots, n$.

Componente	z_i	K_i
Gás carbônico	0,0046	1,650
Metano	0,8345	3,090
Etano	0,0381	0,720
Propano	0,0163	0,390
Isobutano	0,0050	0,210
n-Butano	0,0074	0,175
Pentanos	0,0287	0,093
Hexanos	0,0220	0,065
Heptanos$^+$	0,0434	0,036
	$\sum_{i=1}^{n} z_i = 1$	

Problema 4.14 Em dois reatores tanque de mistura perfeita, de volumes iguais, em série é conduzida uma reação em fase líquida: $nA \longrightarrow$ Produtos, de forma isotérmica. Os balanços estacionários de massa do reagente A neste sistema são descritos pelas equações algébricas:

- Primeiro reator: $kC_1^n = \dfrac{q}{V}[C_0 - C_1]$

- Segundo reator: $k\,C_2^n = \dfrac{q}{V}[C_1 - C_2]$

em que:

k: constante de velocidade da reação $= 0,075\,\dfrac{litro}{mol\ min}$

n: ordem da reação $= 2$

q: vazão volumétrica de alimentação do sistema $= 30\,\dfrac{litro}{min}$

V: volume de cada reator ($litro$)

C_0: concentração do reagente na alimentação do sistema $= 1,6\,\dfrac{mol}{litro}$

C_1: concentração do reagente na saída do primeiro reator $\left[\dfrac{mol}{litro}\right]$

C_0: concentração do reagente na saída do segundo reator $\left[\dfrac{mol}{litro}\right]$

Calcule o volume dos reatores sabendo-se que a conversão global de A é igual a 80 %. Generalize seus resultados para N reatores iguais em série e compare o volume de um reator tubular com escoamento empistonado (PFR) que conduz à mesma conversão. Refaça o problema para reação irreversível de terceira ordem, isto é, $n = 3$.

Sugestão: Considere as seguintes variáveis e parâmetro adimensionais: $y_i = \dfrac{C_i}{C_0}$ para $i = 1, 2, \cdots, n$ e $\alpha = \dfrac{k\,C_0\,V}{q}$.

Problema 4.15 Os balanços de massa de reagente e de energia em um reator de mistura perfeita onde é conduzida uma reação de segunda ordem, irreversível e exotérmica são expressos pelas seguintes equações (em forma adimensional):

- Balanço de massa do reagente: $1 - x = Da\,x^2\ \exp\left[\gamma\left(1 - \dfrac{1}{\theta}\right)\right]$
- Balanço de energia: $\theta - 1 = \beta Da\,x^2\ \exp\left[\gamma\left(1 - \dfrac{1}{\theta}\right)\right]$

Em que: $x = \dfrac{C_{saida}}{C_{entrada}}$ (concentração adimensional do reagente), $\theta = \dfrac{T_{saida}}{T_{atermica}}$ (temperatura adimensional da mistura reacional no interior do reator, sendo $T_{atermica}$ a temperatura no interior do reator se não houvesse geração da calor), Da, β e γ são parâmetros adimensionais cujos valores são $Da = 0,02381$, $\beta = 0,65$ e $\gamma = 20$. Pelo balanço de massa do reagente: $Da\,x^2\ \exp\left[\gamma\left(1 - \dfrac{1}{\theta}\right)\right] = (1 - x)$ que substituído no balanço de energia resulta em: $\theta - 1 = \beta(1 - x) \Rightarrow \theta = 1 + \beta(1 - x)$ que substituída no balanço de massa do reagente dá origem à equação não linear em uma variável:

$$x = 1 - Da\,x^2\ \exp\left[\gamma\left(1 - \frac{1}{1 + \beta(1 - x)}\right)\right] = g(x)$$

A função $g(x)$ é plotada na Figura 4.22 junto com a bissetriz do primeiro quadrante, $y = x$.

Os três pontos de interseção das curvas são:

Ponto	x	θ	
1	0,1822463	1,5315399	alta conversão
2	0,7056739	1,1913120	média conversão
3	0,9656291	1,0223411	baixa conversão

(a) Verifique que quando se aplica o método das substituições sucessivas: $x(k+1) = g(x(k))$ para $k = 0, 1, 2, \cdots$ apenas o ponto 3 é obtido, explique o motivo desta ocorrência;

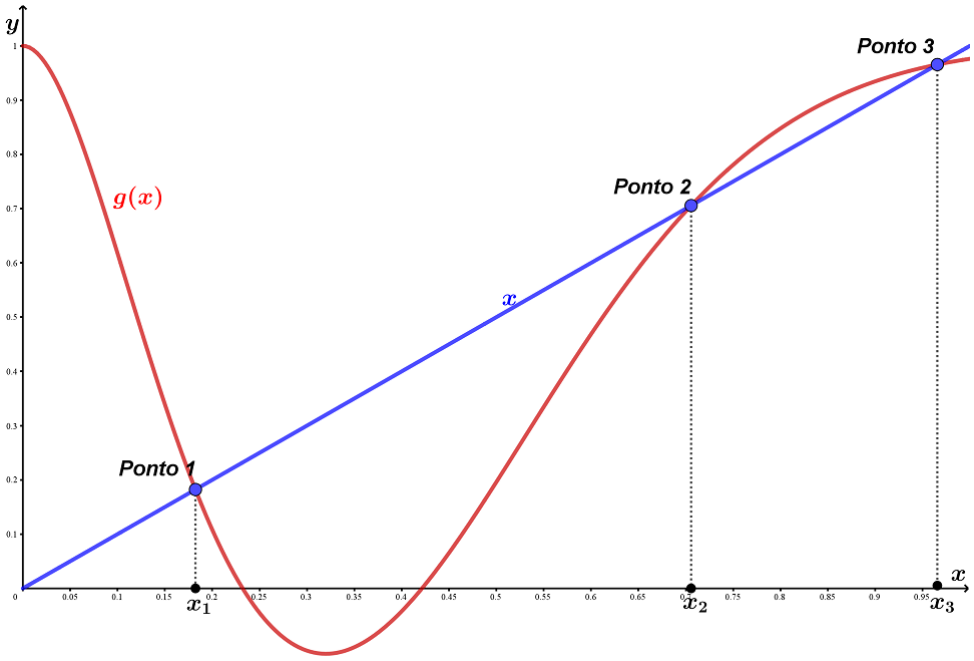


Figura 4.22: Gráfico do Problema 4.15.

(b) Desenvolva um procedimento iterativo que permita determinar todas as soluções e mostre em seu procedimento a condição inicial considerada na obtenção de cada solução.

5. Resolução de Sistemas de Equações Algébricas

5.1 Introdução

Para ilustrar o tipo de problemas abordado no presente capítulo considera-se uma coluna de destilação de três pratos operada de forma contínua na destilação de uma mistura binária, segundo o diagrama da Figura 5.1.

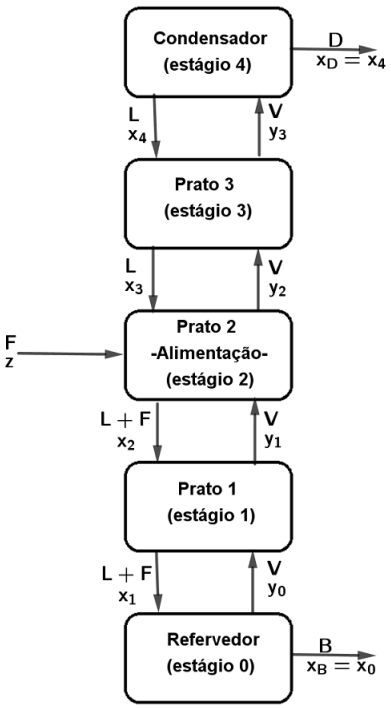


Figura 5.1: Diagrama de um coluna de destilação.

As composições indicadas no diagrama referem-se à fração molar do componente mais leve. A

volatilidade relativa, α, da mistura binária é considerada constante, permitindo expressar:

$$\alpha = \left(\frac{y_i}{1-y_i}\right)\left(\frac{1-x_i}{x_i}\right) \Rightarrow x_i = \frac{y_i}{y_i+\alpha(1-y_i)} \text{ ou } y_i = \frac{\alpha x_i}{\alpha x_i+(1-x_i)}.$$

Os balanços molares do elemento mais volátil em cada um dos estágios são descritos por:

Refervedor (estágio 0): $(1+R)(y_0-x_0)+\left(R+\frac{F}{D}\right)(x_0-x_1)=0$

Prato 1: $(1+R)(y_1-y_0)+\left(R+\frac{F}{D}\right)(x_1-x_2)=0$

Prato 2 (prato de alimentação): $(1+R)(y_2-y_1)+\left(R+\frac{F}{D}\right)x_2-Rx_3=\frac{F}{D}z$

Prato 3: $(1+R)(y_3-y_2)+R(x_3-x_4)=0$

Condensador (estágio 4): $y_3-x_4=0.$

Sendo: $R=\frac{L}{D}$ a chamada *razão de refluxo*, e $V=D+L=(1+R)D$ a vazão molar do vapor.

Além destes balanços têm-se os balanços globais:

$\begin{cases} B+D=F \\ Bx_B+Dx_D=Fz \end{cases}$, sendo $x_B=x_0$ e $x_D=x_4$.

Sendo especificados: $F, z, R,$ e D. As vazões de vapor, de líquido e do fundo podem ser calculadas segundo: $L=RD$, $V=(1+R)D$ e $B=F-D$ restando como incógnitas as composições x_0, x_1, x_2, x_3 e x_4, que após determinadas permitem calcular as composições da fase vapor por $y_i=\frac{\alpha x_i}{\alpha x_i+(1-x_i)}$, para $i=0,1,2,3,4$. O sistema algébrico não linear resultante é:

$$\begin{pmatrix} (1+R)\left(\frac{\alpha x_0}{\alpha x_0+(1-x_0)}-x_0\right)+(R+F/D)(x_0-x_1) \\ (1+R)\left(\frac{\alpha x_1}{\alpha x_1+(1-x_1)}-\frac{\alpha x_0}{\alpha x_0+(1-x_0)}\right)+(R+F/D)(x_1-x_2) \\ (1+R)\left(\frac{\alpha x_2}{\alpha x_2+(1-x_2)}-\frac{\alpha x_1}{\alpha x_1+(1-x_1)}\right)+(R+F/D)x_2-Rx_3 \\ (1+R)\left(\frac{\alpha x_3}{\alpha x_3+(1-x_3)}-\frac{\alpha x_2}{\alpha x_2+(1-x_2)}\right)+R(x_3-x_4) \\ \frac{\alpha x_3}{\alpha x_3+(1-x_3)}-x_4 \end{pmatrix} = \frac{F}{D}z\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}$$

Ou, em notação vetorial: $\mathbf{f}(\mathbf{x})=\frac{F}{D}z\,\mathbf{e_2}$.

Sendo $\mathbf{f}(\mathbf{x}):\mathbb{R}^5\to\mathbb{R}^5$, $\begin{pmatrix}x_0\\x_1\\x_2\\x_3\\x_4\end{pmatrix}\in\mathbb{R}^5$ e $\mathbf{e}_2=\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}$.

Antes de iniciar a busca da solução do problema é importante avaliar se não há *redundância* nas equações utilizadas, por exemplo, se todos os balanços molares forem somados resulta em:

$$\left(\frac{F}{D}-1\right)x_0+x_4=\frac{F}{D}z \Rightarrow (F-D)x_0+Dx_4=Fz,$$

mas $F-D=B$, $x_0=x_B$ e $x_4=x_D$, reproduzindo o balanço global do componente: $Bx_B+Dx_D=Fz$. Dessa forma essa equação está *implicitamente* contida nos balanços dos quatro estágios, não podendo ser utilizada novamente, sua utilização introduziria uma singularidade no sistema.

Configura-se outro problema com a especificação de F, z, x_D e x_B como $x_4=x_D$, e $x_0=x_B$ tem-se agora como as composições x_1, x_2 e x_3, e as vazões molares L, V, D e B. As vazões B e D,

podem ser calculadas pelos balanços globais: $B+D=F$ e $Bx_B+Dx_D=Fz$ e as vazões molares L e V podem ser expressas pela razão de refluxo R: $L=RD$ e $V=(1+R)D$. A composição do vapor no refervedor é calculada por: $y_0=\dfrac{\alpha x_0}{\alpha x_0+(1-x_0)}-x_0$, e pelo balanço no condensador tem-se: $y_3=x_4=x_D \Rightarrow x_3=\dfrac{y_3}{\alpha-(\alpha-1)y_3}$.

Rearranjando as equações de balanço em forma mais adequada, o problema passa a ter apenas três incógnitas x_1, y_2 e R devendo se descartar um dos balanços molares nos estágios, resultando no sistema:

$$\begin{pmatrix} (1+R)(y_0-x_0)+(R+F/D)(x_0-x_1) \\ (1+R)\left(\dfrac{\alpha x_1}{\alpha x_1+(1-x_1)}-y_0\right)+(R+F/D)\left(x_1-\dfrac{y_2}{\alpha-(\alpha-1)y_2}\right) \\ (1+R)(y_3-y_2)+R(x_3-x_4) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Ou, em notação vetorial: $\mathbf{f}(\mathbf{x})=\mathbf{0}$, sendo $\mathbf{f}(\mathbf{x}):\mathbb{R}^3\to\mathbb{R}^3$, $\mathbf{x}=\begin{pmatrix} x_1 \\ y_2 \\ R \end{pmatrix}\in\mathbb{R}^3$ e $\mathbf{0}=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Note que o balanço molar descartado foi o do prato 2 (prato de alimentação), pois esse balanço contém todas as incógnitas do problema e sua não inclusão facilitará a resolução numérica do sistema.

A manipulação desse sistema de equações permite expressar as variáveis x_1 e y_2 em função de R. Assim, da primeira equação obtém-se: $x_1=x_0+\dfrac{(1+R)}{R+F/D}(y_0-x_0)=x_1(R)$ e da terceira equação: $y_2=y_3+\dfrac{R}{1+R}(x_3-x_4)=y_2(R)$. A substituição dessas duas expressões na segunda equação dá origem a:

$$\Phi(R)=(1+R)\left(\frac{\alpha x_1(R)}{\alpha x_1(R)+(1-x_1(R))}-y_0\right)+(R+F/D)\left(x_1(R)-\frac{y_2(R)}{\alpha-(\alpha-1)y_2(R)}\right)=0.$$

Essa última equação pode ser resolvida pelos procedimentos usuais de resolução de equações algébricas não lineares em uma variável (Capítulo 4). Após o valor de R ser determinado calculam-se: as vazões $L=RD$ e $V=(1+R)D$, e as composições x_1 (e y_1) e y_2 (e x_2).

Todo procedimento numérico de resolução de sistemas de equações algébricas não lineares demanda, em cada iteração do procedimento, a resolução de sistemas lineares de mesma dimensão do sistema original. Assim, o desempenho do procedimento é absolutamente dependente da eficiência e acurácia do método de resolução de sistemas lineares de equações algébricas e constitui o principal escopo deste capítulo.

Sistemas algébricos lineares são representados genericamente na forma:

$\mathbf{A}\,\mathbf{x}=\mathbf{b}$, sendo $\mathbf{A}$ matriz quadrada $n\times n$, $\mathbf{x}\in\mathbb{R}^n$: vetor das incógnitas e $\mathbf{b}\in\mathbb{R}^n$: vetor constante.

Existe uma grande variedade de métodos para resolução de sistemas lineares, sendo muitos deles dependentes da estrutura da matriz $\mathbf{A}$ (matriz densa, esparsa, simétrica, tri-diagonal, bloco-diagonal, etc.). Os métodos mais conhecidos para a resolução de sistemas lineares são:

Métodos Diretos: $\begin{cases} \text{Eliminação Gaussiana;} \\ \text{Fatorações } (LU, LL^T, LDL^T, QR, \cdots); \\ \text{Método de Thomas.} \end{cases}$

Métodos Iterativos:
- Método de Jacobi;
- Método de Gauss-Seidel;
- Métodos SOR (*Sucessive Over-Relaxation*);
- Métodos de Minimização.

Sistemas algébricos não lineares são representados genericamente na forma:

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \text{ sendo } \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n \text{ e } \mathbf{x} \in \mathbb{R}^n.$$

Para o caso não linear os seguintes métodos são apresentados:

- Método de Substituições Sucessivas;
- Método de Newton-Raphson;
- Métodos de Broyden;
- Métodos de Minimização;
- Homotopia e Métodos de Continuação.

Como as equações de sistemas de equações algébricas envolvem variáveis de diversas ordens de grandeza é importante, antes de aplicar qualquer procedimento de resolução, realizar um reescalamento das mesmas de modo a mantê-las com ordens de grandeza semelhantes. Um dos métodos de reescalamento é o adimensionamento que, além de reduzir a escala das variáveis, torna o problema independente de sistema de unidades.

Um exemplo do problema de escala em sistemas algébricos não lineares é o sistema hidráulico representado na Figura 5.2.

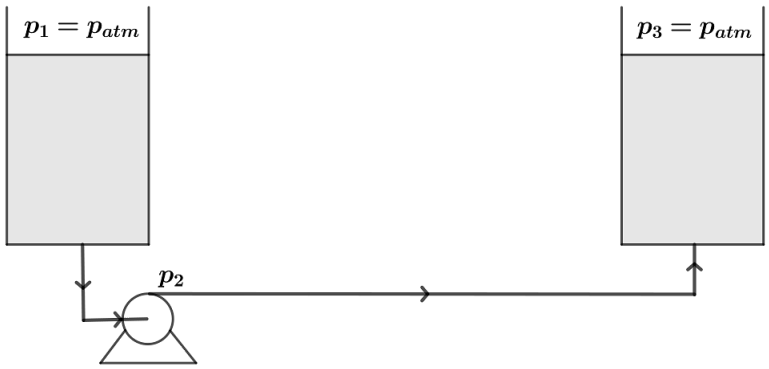


Figura 5.2: Diagrama de um sistema hidráulico.

A bomba centrífuga eleva a pressão do líquido de p_1 (pressão atmosférica) a p_2, mas ocorre uma perda de carga na tubulação que liga os dois tanques a pressão na saída da tubulação cai para p_3 novamente a pressão atmosférica. A elevação da pressão devido à bomba centrífuga é dada por sua curva característica:

$$p_2 - p_1 = a - bQ^{\frac{3}{2}},$$

em que a e b são constantes características da bomba e Q é a vazão volumétrica.

A perda de carga na tubulação é expressa por: $p_2 - p_3 = 8\dfrac{f_M \rho L Q^2}{\pi^2 D^5}$.

Sendo:
- f_M : fator de atrito de Moody (grandeza adimensional);
- ρ; massa específica do líquido;
- L : comprimento da tubulação;
- D : diâmetro interno da tubulação.

Os valores dos parâmetros e grandezas físicas do problema são listadas a seguir:

	Dados 1	Dados 2
D (polegadas)	1,049	2,469
L (pés)	50,0	210,6
f_M (adimensional)	0,032	0,026
$a\,(psi)$	16,7	38,5
$b\left(\frac{psi}{(gpm)^{1,5}}\right)$	0,052	0,0296
$\rho\left(\frac{lb_m}{(ft)^3}\right)$	62,4 (água)	51,4 (querosene)

Deseja-se calcular a pressão p_2 e a vazão Q.

Antes de iniciar a resolução do problema deve-se transformar todos os dados para um mesmo sistema de unidades, no caso o sistema métrico internacional.

	Dados 1	Dados 2
D (metros)	0,027	0,063
L (metros)	15,24	64,191
f_M (adimensional)	0,032	0,026
$a\,(Pa)$	$1,151\times10^5$	$2,654\times10^5$
$b\left(\frac{Pa}{(m^3/s)^{1,5}}\right)$	$7,155\times10^8$	$4,073\times10^8$
$\rho\left(\frac{kg}{m^3}\right)$	999,552 (água)	823,349 (querosene)

$p_1 = p_3 = 1\,atm = 1,013\times10^5\,Pa$.

Como pode-se verificar uma das incógnitas do problema a pressão p_2 é da ordem de grandeza de 10^5 para avaliar a ordem de grandeza da segunda incógnita a vazão Q, deve-se ter uma avaliação aproximada de seu valor. Para isto a equação da bomba é utilizada, calculando-se: $Q_{ref} = \left(\frac{a-(p_2-p_1)}{b}\right)^{\frac{2}{3}}$, considerando $p_2 = 2p_1 = 2\,atm = 2,026\times10^5\,Pa$.

	Dados 1	Dados 2
$Q_{ref}\left(\frac{m^3}{s}\right)$	$0,720\times10^{-3}$	$5,456\times10^{-3}$

Como foi verificado uma grande diferença entre as ordens de grandeza das incógnitas, o seguinte adimensionamento/reescalamento é proposto:

$$\bullet\ \text{Pressão } p_2 \Rightarrow x_1 = \frac{p_2}{p_1} \qquad \bullet\text{Vazão } Q \Rightarrow x_2 = \frac{Q}{Q_{ref}}.$$

Transformando as equações em: $\begin{cases} f_1(x_1,x_2) = x_1 - 1 - \alpha + \beta(x_2)^{\frac{3}{2}}; \\ f_2(x_1,x_2) = x_1 - 1 - \gamma(x_2)^2. \end{cases}$

Sendo: $\alpha = \frac{a}{p_1}$, $\beta = \frac{b(Q_{ref})^{\frac{3}{2}}}{p_1}$ e $\gamma = 8\frac{f_M\,\rho\,L\,(Q_{ref})^2}{\pi^2 D^5 p_1}$, todos adimensionais, correspondendo aos seguintes valores numéricos:

	Dados 1	Dados 2
α	1,136	2,62
β	0,136	1,62
γ	0,150	0,337

Adotando com valor inicial do processo iterativo, para ambos os dados, $\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

Antes dos métodos de resolução de sistema de equações algébricas serem apresentados, na próxima seção são discutidos alguns aspectos relativos a vetores, matrizes e revistos alguns conceitos fundamentais da álgebra linear.

5.2 Análise da Solução de Sistemas Algébricos Lineares

A *norma de um vetor* $\mathbf{x} \in \mathbb{R}^n$ é uma função, $||\cdot||$, de $\mathbb{R}^n \longrightarrow \mathbb{R}$ com as seguintes propriedades:

- $||\mathbf{x}|| \geq 0 \; \forall \mathbf{x} \in \mathbb{R}^n$,
- $||\mathbf{x}|| = 0 \iff \mathbf{x} = \mathbf{0}$,
- $||\alpha \mathbf{x}|| = \alpha ||\mathbf{x}||, \forall \alpha \in \mathbb{R} \wedge \forall \mathbf{x} \in \mathbb{R}^n$,
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||, \forall \mathbf{x} \wedge \forall \mathbf{y} \in \mathbb{R}^n$.

As normas mais comuns são:

1. Norma máxima (l_∞) : $||\mathbf{x}||_\infty = \arg\max_{i=1}^{n} |x_i|$,
2. Norma absoluta (l_1) : $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$,
3. Norma euclidiana (l_2) : $||\mathbf{x}||_2 = \sum_{i=1}^{n} \sqrt{|x_i|^2}$.

No presente texto, apenas a norma euclidiana de vetores é considerada, dispensando a adição do subscrito em sua representação. Além disto, $\forall \mathbf{x} \in \mathbb{R}^n$ é representado na forma de um *vetor coluna*: $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$. Um *vetor linha* é representado por: $\mathbf{x}^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$.

O *produto escalar* de dois vetores $\mathbf{x} \wedge \mathbf{y} \in \mathbb{R}^n$, representado por $\mathbf{x} \cdot \mathbf{y}$ ou $\mathbf{x}^T \mathbf{y}$, é definido por: $\sum_{i=1}^{n} x_i \, y_i$, verificando-se a comutatividade desta operação, assim:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x} = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_{i=1}^{n} x_i \, y_i.$$

Os vetores $\mathbf{x} \wedge \mathbf{y} \in \mathbb{R}^n$ são ditos *ortogonais* se o produto escalar $\mathbf{x} \cdot \mathbf{y} = 0$.

Da definição do produto escalar, conclui-se que:

$$\mathbf{x} \cdot \mathbf{x} = \mathbf{x}^T \mathbf{x} = \sum_{i=1}^{n} x_i^2 = ||\mathbf{x}||^2.$$

$m \leq n$ vetores são ditos linearmente independentes se: $c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_m \mathbf{x}_m = \mathbf{0}$ se e somente se $c_1 = c_2 = \cdots = c_m = 0$.

A dimensão n de um *espaço vetorial* é o número máximo de vetores linearmente independentes que esse espaço admite. Assim, se $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ são n vetores linearmente independentes no $\mathbb{R}^n$, os mesmos podem ser considerados como uma *base* do $\mathbb{R}^n$ e qualquer outro vetor no $\mathbb{R}^n$

pode ser expresso por uma combinação linear desses vetores, segundo: $\hat{\mathbf{x}} = \sum_{i=1}^{n} a_i \, \mathbf{x}_i$, as constantes a_i (para $i = 1, 2, \cdots, n$) são chamados de *componentes* do vetor $\hat{\mathbf{x}}$ nessa base. Se além de linearmente independentes, esses vetores apresentarem a norma unitária e forem mutuamente ortogonais, isto é: $\mathbf{x}_i \cdot \mathbf{x}_j = \delta_{i,j}$, a base é dita *ortonormal*. A base ortonormal padrão do $\mathbb{R}^n$ é a *base canônica* designada por $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$ em que o componente i do vetor $\mathbf{e}_j$ é igual a $\delta_{i,j}$. Os componentes de qualquer vetor $\mathbf{x} \in \mathbb{R}^n$ na base canônica são seus elementos e $x_i = \mathbf{x} \cdot \mathbf{e}_i$.

A forma genérica de representação de um sistema algébrico linear $\mathbf{A} \cdot \mathbf{x} = \mathbf{z}$ em que $\mathbf{A}$ é uma matriz retangular com m linhas e n colunas, $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{z} \in \mathbb{R}^m$, expresso na forma indicial por $z_i = \sum_{j=1}^{n} a_{i,j} \, x_j$ para $i = 1, 2 \cdots, m$, pode ser considerada com uma transformação linear de um vetor $\mathbf{x} \in \mathbb{R}^n$ em um vetor $\mathbf{z} \in \mathbb{R}^m$. Representando a matriz $\mathbf{A}$ por seus vetores coluna, isto é $\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}$, sendo $\mathbf{a}_i \in \mathbb{R}^m$, para $i = 1, 2, \cdots, n$, pode se representar $\mathbf{A} \cdot \mathbf{x} = \sum_{i=1}^{n} x_i \, \mathbf{a}_i = \mathbf{z}$, então $\mathbf{z}$ é uma combinação linear dos n vetores coluna da matriz $\mathbf{A}$, implicando que a matriz $\mathbf{A}_{aumentada} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n & \mathbf{z} \end{pmatrix}$ deve ter o mesmo número de vetores coluna linearmente independentes que o da matriz original $\mathbf{A}$, para o sistema linear poder ser classificado como *consistente*. O número de vetores coluna (ou vetores linha) linearmente independentes que uma matriz contém é o chamado *posto* da matriz, assim o sistema $\mathbf{A} \cdot \mathbf{x} = \mathbf{z}$ só é consistente se o posto de $\mathbf{A}$ for igual ao posto da matriz $\mathbf{A}_{aumentada} = [\mathbf{A} \,|\, \mathbf{z}]$. Três situações podem ocorrer:

(a) Número de equações menor que o número de incógnitas, isto é $m < n$. Neste caso, m é o valor máximo do posto da matriz $\mathbf{A}$, sendo r o valor do posto, necessariamente deve-se ter $r \leq m$ e, para que o sistema seja consistente, o posto de $\mathbf{A}_{aumentada}$ também deve ser igual a r. Neste caso, se o sistema for consistente pode-se afirmar que apresenta um número infinito de soluções. Na realidade, isto evidencia que r incógnitas podem ser expressas por uma combinação linear de $(n-r)$ incógnitas.

(b) Número de equações igual o número de incógnitas, isto é $m = n$. A matriz $\mathbf{A}$ é uma *matriz quadrada*, $\mathbf{x} \in \mathbb{R}^n$ é o *vetor de incógnitas* e $\mathbf{z} = \mathbf{b} \in \mathbb{R}^n$ é chamado de *vetor das constantes*, com componentes conhecidos. Este problema só apresenta solução se a matriz $\mathbf{A}$ for *regular*, isto é se existir uma matriz $\mathbf{A}^{-1}$, chamada de *matriz inversa*, tal que $\mathbf{A}^{-1} \, \mathbf{A} = \mathbf{A} \, \mathbf{A}^{-1} = \mathbf{I}$ em que $\mathbf{I}$ é a *matriz identidade*. E a solução do sistema é: $\mathbf{x} = \mathbf{A}^{-1} \, \mathbf{b}$. Assim, a matriz $\mathbf{A}\,(n \times n)$ é regular se tiver o posto igual a n, a condição necessária e suficiente para que tal ocorra é $\det(\mathbf{A}) \neq 0$, não havendo possibilidade de o posto da matriz $\mathbf{A}_{aumentada} = [\mathbf{A} \,|\, \mathbf{b}]$, ser maior do que n. Se o sistema for consistente, porém o posto de **A** = posto de $\mathbf{A}_{aumentada} = r < n$, a situação é semelhante à ocorrida no caso (a).

(c) Número de equações é maior que o número de incógnitas. Como neste caso a matriz $\mathbf{A}$ apresenta um número de colunas, n, menor do que o número de linhas, m, o valor máximo do posto de $\mathbf{A}$ é n, como os vetores coluna de $\mathbf{A}$ são de dimensão $m > n$ existe a possibilidade do posto de $\mathbf{A}_{aumentada}$ ser igual a $(n+1)$, valor maior que o posto de $\mathbf{A}$ tornando o sistema inconsistente. A consistência do sistema nesta situação é um indicativo de que as equações que estão em excesso, $(m-n)$ equações, não contradizem as n equações consistentes utilizadas na resolução do sistema e a solução do sistema existe e é única. Havendo consistência do sistema, mas o posto de **A** = posto de $\mathbf{A}_{aumentada} = r < n$ volta a ocorrer a situação apresentada nos dois casos anteriores.

Esta análise pode ser sintetizada no diagrama da Figura 5.3.

Para ilustrar o procedimento, apresenta-se a análise de consistência dos seguintes sistemas lineares de baixa dimensão:

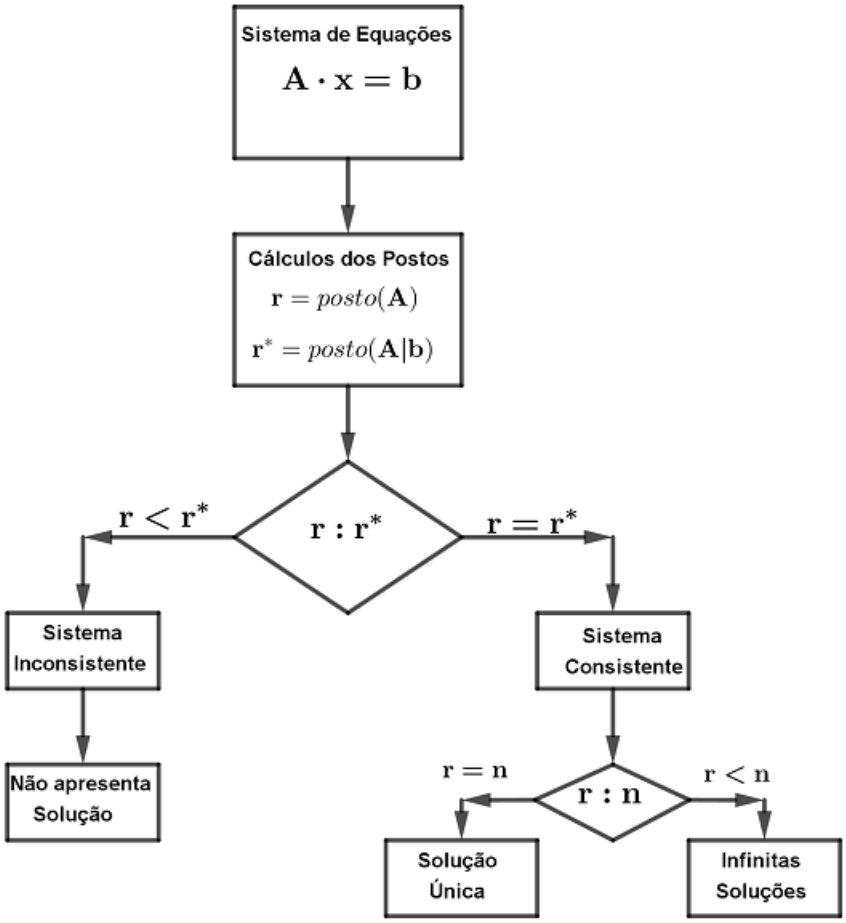


Figura 5.3: Análise de soluções de sistemas lineares.

- Três equações ($m = 3$) e duas incógnitas ($n = 2$).

$$\begin{cases} x_1 + x_2 = 3 \\ 2x_1 + 2x_2 = 0 \\ x_1 - x_2 = -1 \end{cases} \Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 1 & -1 \end{pmatrix}, r = posto(\mathbf{A}) = 2;$$

$$\mathbf{A}_{aumentada} = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 0 \\ 1 & -1 & -1 \end{pmatrix}, r^* = posto(\mathbf{A}_{aumentada}) = 3 \Rightarrow r^* > r \text{ sistema inconsistente.}$$

Por ser um sistema de baixa dimensão, verifica-se facilmente que o lado esquerdo da segunda equação é o dobro do lado esquerdo da primeira e, para as que as duas equações não sejam conflitantes, o mesmo deveria ocorrer com os lados direitos de ambas, o que não ocorre.

- Duas equações ($m = 3$) e duas incógnitas ($n = 2$).

$$\begin{cases} 2x_1 + 3x_2 = 0 \\ x_1 + x_2 = 0 \end{cases} \Rightarrow \mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}, r = posto(\mathbf{A}) = 2;$$

$$\mathbf{A}_{aumentada} = \begin{pmatrix} 2 & 3 & 0 \\ 1 & 1 & 0 \end{pmatrix}, r^* = posto(\mathbf{A}_{aumentada}) = 2 \Rightarrow r^* = r \text{ sistema consistente,}$$

e apresenta solução única (solução trivial) $x_1 = x_2 = 0$. Reforçando o já exposto que todo sistema homogêneo em que o posto de $\mathbf{A}$ é igual a n apresenta com única solução a solução trivial.

- Duas equações ($m = 3$) e três incógnitas ($n = 3$).

$$\begin{cases} x_1 + 2x_2 + x_3 = 4 \\ 3x_1 + 6x_2 - x_3 = 8 \end{cases} \Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 6 & -1 \end{pmatrix}, r = posto(\mathbf{A}) = 2;$$

$$\mathbf{A}_{aumentada} = \begin{pmatrix} 1 & 2 & 1 & 4 \\ 3 & 6 & -1 & 8 \end{pmatrix}, r^* = posto(\mathbf{A}_{aumentada}) = 2 \Rightarrow r^* = r,$$

sistema consistente (o que sempre ocorre quando $m < n$ e $posto(\mathbf{A}) = m$ e indicando que m vetores coluna de $\mathbf{A}$ constituem uma base de $\mathbb{R}^m$. Dessa forma o vetor $\mathbf{b}$ necessariamente é linearmente dependente desses vetores e, em consequência, a matriz $\mathbf{A}_{aumentada}$ apresenta sempre posto igual ao de $\mathbf{A}$, além disto como $r = m < n$ o sistema, neste caso, apresenta um número infinito de soluções). Entretanto, as soluções deste problema apresentam uma restrição indicada pela combinação linear das duas primeiras colunas da matriz $\mathbf{A}$, observando que o sistema pode ser reescrito na forma:

$$\begin{cases} (x_1 + 2x_2) + x_3 = z + x_3 = 4 \\ 3(x_1 + 2x_2) - x_3 = 3z - x_3 = 8 \end{cases},$$

sendo $z = x_1 + 2x_2 \Rightarrow \begin{cases} z = 3 \\ x_3 = 1 \end{cases}$.

Então, o sistema de fato apresenta um número infinito de soluções, porém o único valor admissível para a incógnita x_3 é o unitário e as variáveis x_1 e x_2 estão sobre a reta $x_1 + 2x_2 = 3$.

Se $\hat{\mathbf{x}}$ for uma solução computada para o sistema linear $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, então seu **erro** é a diferença $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$. É claro que este erro não é conhecido, pois se o fosse a solução exata ($\mathbf{x}$) já seria conhecida, dispensando discussões adicionais. Entretanto, uma maneira de avaliar a qualidade da solução obtida $\hat{\mathbf{x}}$ é calculando o **resíduo** definido por $\mathbf{r} = \mathbf{A}\,\mathbf{x} - \mathbf{A}\,\hat{\mathbf{x}}$ como $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, então $\mathbf{r} = \mathbf{b} - \mathbf{A}\,\hat{\mathbf{x}}$. Este resíduo *mede* o grau de satisfação de $\hat{\mathbf{x}}$ no atendimento da restrição $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Se $\mathbf{r}$ for igual a zero, então $\hat{\mathbf{x}}$ é a solução exata e o erro $\mathbf{e}$ é nulo. Se $\hat{\mathbf{x}}$ for uma boa aproximação da solução exata, espera-se que cada elemento de $\mathbf{r}$ seja próximo de zero.

Para ilustrar esses dois conceitos o sistema linear $\begin{cases} 2{,}02x_1 + 1{,}98x_2 = 3 \\ 1{,}98x_1 + 2{,}02x_2 = 3 \end{cases}$ é considerado, cuja solução exata é $\mathbf{x} = \begin{pmatrix} 0{,}75 \\ 0{,}75 \end{pmatrix}$.

Caso $\hat{\mathbf{x}} = \begin{pmatrix} 0{,}7575 \\ 0{,}7575 \end{pmatrix} \Rightarrow \mathbf{e} = -\begin{pmatrix} 0{,}0075 \\ 0{,}0075 \end{pmatrix}$ e $\mathbf{r} = -\begin{pmatrix} 0{,}003 \\ 0{,}003 \end{pmatrix}$, neste caso tanto o erro (da ordem de 1% em relação à $\mathbf{x}$) quanto o resíduo (da ordem de 1% em relação à $\mathbf{b}$) são pequenos.

Caso $\hat{\mathbf{x}} = \begin{pmatrix} 1{,}5 \\ 0 \end{pmatrix} \Rightarrow \mathbf{e} = \begin{pmatrix} -0{,}75 \\ 0{,}75 \end{pmatrix}$ e $\mathbf{r} = \begin{pmatrix} -0{,}003 \\ 0{,}003 \end{pmatrix}$, neste caso o erro (da ordem de 100% em relação à $\mathbf{x}$) é elevado, porém o resíduo (da ordem de 1% em relação à $\mathbf{b}$) é pequeno.

Para o sistema linear $\begin{cases} 2{,}02x_1 + 1{,}98x_2 = 3 \\ 1{,}98x_1 + 2{,}02x_2 = -3 \end{cases}$, cuja solução exata é $\mathbf{x} = \begin{pmatrix} 75 \\ -75 \end{pmatrix}$

Caso $\hat{\mathbf{x}} = \begin{pmatrix} 76 \\ -74 \end{pmatrix} \Rightarrow \mathbf{e} = -\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ e $\mathbf{r} = -\begin{pmatrix} 4 \\ 4 \end{pmatrix}$, neste caso o erro (da ordem de 1,3% em relação à $\mathbf{x}$) é pequeno, porém o resíduo (da ordem de -133% em relação à $\mathbf{b}$) é elevado.

Além da análise de consistência de sistemas algébrico lineares é aconselhável também avaliar o dito *condicionamento* da matriz $\mathbf{A}$ do sistema, considerada quadrada e não singular. Uma matriz $\mathbf{A}$ é considerada *mal condicionada* quando apresenta um *número* ou *grau de condicionamento* elevado. Valores elevados dos números de condicionamento são indicativos de dificuldades numéricas na resolução do sistema e na inversão da matriz $\mathbf{A}$. Os quatro números de condicionamento mais usuais são:

(1) $\mathbf{M} = \mathbf{M}(\mathbf{A})\,\mathbf{M}(\mathbf{A}^{-1})$ em que $\mathbf{M}(\mathbf{A}) = \max\limits_{i} \sum\limits_{j=1}^{n} |a_{ij}|$ que é o valor da soma dos valores absolutos dos elemento da linha i da matriz $\mathbf{A}$ que apresenta o maior valor.

(2) $\mathbf{N} = \mathbf{N}(\mathbf{A})\,\mathbf{N}(\mathbf{A}^{-1})$ em que $\mathbf{N}(\mathbf{A}) = \sqrt{\text{tr}(\mathbf{A}\,\mathbf{A}^T)}$ norma euclidiana de $\mathbf{A}$.

(3) $\mathbf{P} = \dfrac{|\lambda|}{|\mu|}$ em que $|\lambda|$ e $|\mu|$ são, respectivamente os valores absolutos do maior e do menor valor característico de $\mathbf{A}$ em módulo (ou da parte real dos mesmos).

(4) $\kappa = \frac{\sigma_{max}(\mathbf{A})}{\sigma_{min}(\mathbf{A})}$ em que $\sigma(\mathbf{A})$ são os valores singulares de $\mathbf{A}$ ou a raiz quadrada dos valores característicos de $\mathbf{A}\,\mathbf{A}^T$.

■ **Exemplo 5.1** Um exemplo do problema de condicionamento de sistemas algébricos lineares é o problema de T.S. Wilson (Marcus, 1960).

$$\begin{cases} 10x_1 + 7x_2 + 8x_3 + 7x_4 = 32 \\ 7x_1 + 5x_2 + 6x_3 + 5x_4 = 23 \\ 8x_1 + 6x_2 + 10x_3 + 9x_4 = 33 \\ 7x_1 + 5x_2 + 9x_3 + 10x_4 = 31 \end{cases}, \text{ cuja solução é } \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Com $\hat{\mathbf{x}} = \begin{pmatrix} 6 \\ -7,2 \\ 2,9 \\ -0,1 \end{pmatrix} \Rightarrow \mathbf{e} = \begin{pmatrix} -5 \\ 8,2 \\ -1,9 \\ 1,1 \end{pmatrix}$ e $\mathbf{r} = \begin{pmatrix} -0,1 \\ 0,1 \\ 0,1 \\ -0,1 \end{pmatrix}$ neste caso o erro (o módulo do erro máximo é da ordem de 820% em relação à $\mathbf{x}$) é elevado, porém o resíduo (o módulo do resíduo máximo é da ordem de 0,435% em relação à $\mathbf{b}$) é pequeno.

Com $\hat{\mathbf{x}} = \begin{pmatrix} 1,5 \\ 0,18 \\ 1,19 \\ 0,89 \end{pmatrix} \Rightarrow \mathbf{e} = \begin{pmatrix} -0,5 \\ 0,82 \\ -0,19 \\ 0,11 \end{pmatrix}$ e $\mathbf{r} = \begin{pmatrix} -0,01 \\ 0,01 \\ 0,01 \\ -0,01 \end{pmatrix}$ neste caso o erro (o módulo do erro máximo é da ordem de 82% em relação à $\mathbf{x}$) ainda é elevado e o resíduo (o módulo do resíduo máximo é da ordem de 0,0435% em relação à $\mathbf{b}$) mantém-se pequeno.

Os quatro números de condicionamento do problema são $\begin{cases} \mathbf{M} = 4488 \\ \mathbf{N} = 3009,579 \\ \mathbf{P} = 2984,093 \\ \kappa = 2984,093 \end{cases}$, tais valores elevados indicam o mau condicionamento do sistema. Os últimos dois números de condicionamento são iguais para matrizes simétrica. ■

5.3 Pivotamento e Método de Eliminação de Gauss

A finalidade do *Método de Eliminação de Gauss*[1] é reduzir a matriz $\mathbf{A}$ a uma estrutura triangular (método de triangularização) ou diagonal (*Método de Eliminação de Gauss-Jordan*[2]) através de

[1] Johann Carl Friedrich Gauss (1777-1855).
[2] Wilhelm Jordan (1842-1899).

operações da álgebra elementar.

O método de diagonalização pode ser implementado pelo algoritmo.

Entre com n (dimensão do sistema)
$\mathbf{A} = (a_{i,j})$ e $\mathbf{b} = (b_i)$ com $i, j = 1, 2, \cdots, n$.
Construa a matriz $\mathbf{A}_{aumentada} = [\mathbf{A} \,|\, \mathbf{b}]$

Para $i = 1, 2, \cdots, n,$ faça
  $\alpha \longleftarrow a_{i,i}$
  Para $j = 1, 2, \cdots, n+1,$ faça
    $a_{i,j} \longleftarrow \dfrac{a_{i,j}}{\alpha}$
  Para $k = 1, 2, \cdots, n \wedge k \neq i,$ faça
    $\alpha \longleftarrow a_{k,i}$
    Para $j = 1, 2, \cdots, n+1,$ faça
      $a_{k,j} \longleftarrow a_{k,j} - \alpha\, a_{i,j}$

No final tem-se na matriz $\mathbf{A}_{aumentada} = [\mathbf{I} \,|\, \mathbf{x}]$.

O método de triangularização pode ser implementado pelo algoritmo.

Entre com n (dimensão do sistema)
$\mathbf{A} = (a_{i,j})$ e $\mathbf{b} = (b_i)$ com $i, j = 1, 2, \cdots, n$.
Construa a matriz $\mathbf{A}_{aumentada} = [\mathbf{A} \,|\, \mathbf{b}]$

Para $i = 1, 2, \cdots, n,$ faça
  $\alpha \longleftarrow a_{i,i}$
  Para $j = 1, 2, \cdots, n+1,$ faça
    $a_{i,j} \longleftarrow \dfrac{a_{i,j}}{\alpha}$
  Para $k = i+1 \cdots, n \wedge i < n,$ faça
    $\alpha \longleftarrow a_{k,i}$
    Para $j = 1, 2, \cdots, n+1,$ faça
      $a_{k,j} \longleftarrow a_{k,j} - \alpha\, a_{i,j}$

$x_n \longleftarrow a_{n,n+1}$

Para $i = n-1, n-2, \cdots, 0,$ faça
  $x_i \longleftarrow a_{i,n+1}$
  Para $j = i+1, \cdots, n,$ faça
    $x_i \longleftarrow x_i - a_{i,j}\, x_j$

No final tem-se a solução do sistema $\mathbf{x}$ e na matriz $\mathbf{A}_{aumentada} = [\mathbf{U} \,|\, \hat{\mathbf{b}}]$, sendo $\mathbf{U}$ uma matriz *triangular superior* da forma:

$$\mathbf{U} = \begin{pmatrix} 1 & u_{1,2} & u_{1,3} & \cdots & u_{1,n-1} & u_{1,n} \\ 0 & 1 & u_{2,3} & \cdots & u_{2,n-1} & u_{2,n} \\ 0 & 0 & 1 & \cdots & u_{3,n-1} & u_{3,n} \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & u_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

De modo a evitar prováveis divisões por zero (dos elementos $a_{i,i}$) e para garantir também a estabilidade numérica do algoritmo (devido a problemas de arredondamento), faz-se necessário o uso de técnicas de **pivotamento**. Pivotamentos são operações de trocas de linhas e/ou colunas de modo a obter uma matriz tendo na diagonal elementos com maior valor absoluto. Quando são

efetuadas somente trocas de linhas, diz-se um pivotamento parcial. No pivotamento total tem-se trocas de linhas e colunas.

Antes dos dois processos de eliminação (diagonalização e triangularização) o procedimento de pivotamento deve ser aplicado, evitando assim a possibilidade de termos da diagonal serem nulos (ou próximo de zero). Uma forma de pivotamento total pode ser implementada pelo seguinte algoritmo.

Entre com n (dimensão do sistema)
$\mathbf{A} = (a_{i,j})$ e $\mathbf{b} = (b_i)$ com $i, j = 1, 2, \cdots, n$.
Construa a matriz $\mathbf{A}_{aumentada} = [\mathbf{A} \,|\, \mathbf{b}]$
$J \longleftarrow 0$
flag $\longleftarrow 1$
$\mathbf{II} \longleftarrow$ Matriz identidade (n, n)
$\mathbf{P} \longleftarrow \mathbf{II}$
$\mathbf{Q} \longleftarrow \mathbf{II}$
$\mathbf{p} \longleftarrow \mathbf{II}$
$\mathbf{q} \longleftarrow \mathbf{II}$

Enquanto flag $\neq 0$, faça
$K \longleftarrow 0$
Para $i = 1, 2, \cdots, n$ faça
$M \longleftarrow |a_{i,i}|$
$k \longleftarrow i$
Se $i < n$ faça para $j = i+1, \cdots, n$
Se $|a_{j,i}| > M$ então
$k \longleftarrow j$
$M \longleftarrow |a_{j,i}|$
Se $k \neq i$ então
$\mathbf{p}^{<k>} \longleftarrow \mathbf{II}^{<i>}$
$\mathbf{p}^{<i>} \longleftarrow \mathbf{II}^{<k>}$
faça para $j = 1, \cdots, n+1$
$c_j \longleftarrow a_{k,j}$
$a_{k,j} \longleftarrow a_{i,j}$
$a_{i,j} \longleftarrow c_j$
$\mathbf{P} \longleftarrow \mathbf{p}\,\mathbf{P}$
$\mathbf{p} \longleftarrow \mathbf{II}$
Se $|a_{i,i}| > TOL$ então $K \longleftarrow K + 1$
Se $K = n \vee J = n$ então flag $\longleftarrow 0$
Senão, faça
$J \longleftarrow J + 1$
Para $i = 1, 2, \cdots, n$
Se $|a_{i,i}| < TOL$ então $k \longleftarrow i$
Para $i = 1, 2, \cdots, n$ faça
$c_i \longleftarrow a_{i,k}$
$a_{i,k} \longleftarrow a_{i,1}$
$a_{i,1} \longleftarrow c_i$
$\mathbf{q}^{<k>} \longleftarrow \mathbf{II}^{<1>}$
$\mathbf{q}^{<1>} \longleftarrow \mathbf{II}^{<k>}$
$\mathbf{Q} \longleftarrow \mathbf{Q}\,\mathbf{q}$
$\mathbf{q} \longleftarrow \mathbf{II}$

Saídas: Nova $\mathbf{A}_{aumentada}$, $\mathbf{P}$ e $\mathbf{Q}$.

Antes de ser mostrado um exemplo envolvendo os procedimentos de eliminação e de pivotamento de matrizes, apresentam-se métodos matriciais que promovem trocas de linhas e de colunas em sistemas algébricos lineares. Para ambas as operações são utilizadas formas modificadas da matriz identidade, assim para trocar a linha i com a linha j de uma matriz quadrada $\mathbf{A}$ a mesma é pré-multiplicada pela forma modificada de matriz identidade em que a linha i é trocada com a linha j, por exemplo, se $n = 4$, $i = 3$ e $j = 1$, a forma modificada da matriz identidade seria: $\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ verificando-se que: $\mathbf{P} \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{pmatrix} = \begin{pmatrix} a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{pmatrix}$.

Para trocar a coluna i com a coluna j de uma matriz quadrada $\mathbf{A}$ a mesma é pós-multiplicada pela forma modificada de matriz identidade em que a coluna i é trocada com a coluna j, como a matriz identidade é uma matriz simétrica com todos os elementos não nulos iguais à unidade, esta troca é igual à troca da linha i é pela linha j. Por exemplo, se $n = 4$, $i = 3$ e $j = 1$, a forma modificada da matriz identidade é igual à matriz $\mathbf{P}$ acima e:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{pmatrix} \mathbf{P} = \begin{pmatrix} a_{1,3} & a_{1,2} & a_{1,1} & a_{1,4} \\ a_{2,3} & a_{2,2} & a_{2,1} & a_{2,4} \\ a_{3,3} & a_{3,2} & a_{3,1} & a_{3,4} \\ a_{4,3} & a_{4,2} & a_{4,1} & a_{4,4} \end{pmatrix}.$$

Matrizes da forma da matriz $\mathbf{P}$ são também iguais à sua inversa, pois $\mathbf{P}\,\mathbf{P} = \mathbf{I}$ porque a multiplicação de $\mathbf{P}$ pela própria matriz $\mathbf{P}$ *desfaz* a troca da linha (ou coluna) feita na matriz identidade, implicando no retorno à sua forma original.

A pré-multiplicação de uma matriz do tipo de $\mathbf{P}$ a ambos os membros do sistema algébrico linear: $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ resulta em: $(\mathbf{P}\,\mathbf{A})\mathbf{x} = (\mathbf{P}\,\mathbf{b})$ transformando $\mathbf{A}^* = \mathbf{P}\,\mathbf{A}$ e $\mathbf{b}^* = \mathbf{P}\,\mathbf{b}$ e o sistema em: $\mathbf{A}^*\,\mathbf{x} = \mathbf{b}^*$ cuja solução $\mathbf{x}$ é a mesma do sistema original. Na realidade esta multiplicação é equivalente a trocar a ordem das equações do sistema, o que em nada afeta o valor das incógnitas. Já a troca de colunas da matriz $\mathbf{A}$ afeta também a ordem do vetor das incógnitas, assim se $\mathbf{Q}$ é uma matriz semelhante à matriz $\mathbf{P}$ (em que há apenas uma troca de linhas da matriz identidade) e, como $\mathbf{Q}\,\mathbf{Q} = \mathbf{I}$, resulta que $\mathbf{A}\,\mathbf{x} = (\mathbf{A}\,\mathbf{Q})\,(\mathbf{Q}\,\mathbf{x}) = \mathbf{b}$ identificando $\hat{\mathbf{A}} = (\mathbf{A}\,\mathbf{Q})$ e $\hat{\mathbf{x}} = (\mathbf{Q}\,\mathbf{x})$ e o sistema algébrico linear transforma-se em: $\hat{\mathbf{A}}\,\hat{\mathbf{x}} = \mathbf{b}$, neste caso o vetor de constantes $\mathbf{b}$ não é afetado e no vetor das incógnitas o elemento i é trocado pelo elemento j (a mesma troca verificada entre a coluna i e a coluna j da matriz $\mathbf{A}$).

■ **Exemplo 5.2** Ilustração do Método de Eliminação de Gauss.

Resolução do Sistema Linear: $\begin{cases} 10x_1 + 2x_2 + 3x_3 + 5x_4 = 43 \\ 2x_1 + 5x_2 + 6x_3 - 2x_4 = 22 \\ 12x_1 + 2x_2 + x_4 = 20 \\ 10x_1 + x_2 = 12 \end{cases}$

Permitindo identificar: $\mathbf{A} = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 2 & 5 & 6 & -2 \\ 12 & 2 & 0 & 1 \\ 10 & 1 & 0 & 0 \end{pmatrix}$ e $\mathbf{b} = \begin{pmatrix} 43 \\ 22 \\ 20 \\ 12 \end{pmatrix}$, assim a matriz $\mathbf{A}$ *aumentada* é:

$$\mathbf{A}_{aumentada} = \begin{pmatrix} 10 & 2 & 3 & 5 & 43 \\ 2 & 5 & 6 & -2 & 22 \\ 12 & 2 & 0 & 1 & 20 \\ 10 & 1 & 0 & 0 & 12 \end{pmatrix}.$$

Antes de aplicar os procedimentos de eliminação, o sistema é submetido a um pivotamento.

Analisando a primeira coluna da matriz $\mathbf{A}_{aumentada}$ identifica-se o maior elemento (em módulo)

como o elemento da terceira linha, é feita então a troca da primeira com a terceira linha.

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 10 & 2 & 3 & 5 & 43 \\ 2 & 5 & 6 & -2 & 22 \\ 12 & 2 & 0 & 1 & 20 \\ 10 & 1 & 0 & 0 & 20 \end{pmatrix} = \begin{pmatrix} 12 & 2 & 0 & 1 & 20 \\ 2 & 5 & 6 & -2 & 22 \\ 10 & 2 & 3 & 5 & 43 \\ 10 & 1 & 0 & 0 & 12 \end{pmatrix}.$$

Não há troca de linhas pela análise da segunda coluna (da segunda linha para baixo), também não há trocas pela análise da terceira coluna. Na quarta coluna aparece um elemento nulo na diagonal, tornando necessária a troca da primeira coluna com a quarta coluna. Assim:

$$\begin{pmatrix} 12 & 2 & 0 & 1 \\ 2 & 5 & 6 & -2 \\ 10 & 2 & 3 & 5 \\ 10 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 & 12 \\ -2 & 5 & 6 & 2 \\ 5 & 2 & 3 & 10 \\ 0 & 1 & 0 & 10 \end{pmatrix} \text{ e } \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_4 \\ x_2 \\ x_3 \\ x_1 \end{pmatrix}$$

A matriz $\mathbf{A}_{aumentada}$ é então:

$$\begin{pmatrix} 1 & 2 & 0 & 12 & 20 \\ -2 & 5 & 6 & 2 & 22 \\ 5 & 2 & 3 & 10 & 43 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}$$

Analisando a primeira coluna o maior elemento (em módulo) é o da terceira linha, é feita então a troca da primeira com a terceira linha.

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 12 & 20 \\ -2 & 5 & 6 & 2 & 22 \\ 5 & 2 & 3 & 10 & 43 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 3 & 10 & 43 \\ -2 & 5 & 6 & 2 & 22 \\ 1 & 2 & 0 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}.$$

Não há troca de linhas com a análise da segunda coluna (da segunda linha para baixo), como os dois elementos da terceira coluna na terceira e quarta linhas são nulos deve-se trocar esta coluna com a primeira. Assim:

$$\begin{pmatrix} 5 & 2 & 3 & 10 \\ -2 & 5 & 6 & 2 \\ 1 & 2 & 0 & 12 \\ 0 & 1 & 0 & 10 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 5 & 10 \\ 6 & 5 & -2 & 2 \\ 0 & 2 & 1 & 12 \\ 0 & 1 & 0 & 10 \end{pmatrix} \text{ e } \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_4 \\ x_2 \\ x_3 \\ x_1 \end{pmatrix} = \begin{pmatrix} x_3 \\ x_2 \\ x_4 \\ x_1 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 2 & 5 & 10 & 43 \\ 6 & 5 & -2 & 2 & 22 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}.$$

Analisando a primeira coluna o maior elemento (em módulo) é o da segunda linha, é feita então a troca da primeira com a segunda linha.

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & 5 & 10 & 43 \\ 6 & 5 & -2 & 2 & 22 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix} = \begin{pmatrix} 6 & 5 & -2 & 2 & 22 \\ 3 & 2 & 5 & 10 & 43 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix} \text{ e } \hat{\mathbf{x}} = \begin{pmatrix} x_3 \\ x_2 \\ x_4 \\ x_1 \end{pmatrix}.$$

Como não há elementos nulos na diagonal da última forma da matriz **A**, não há mais necessidade de trocas de linhas ou de colunas, concluindo-se assim a fase de pivotamento do sistema. Para transformar o sistema nesta forma final, aplicaram-se as seguintes transformações à matriz **A**.

$$\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}10&2&3&5\\2&5&6&-2\\12&2&0&1\\10&1&0&0\end{pmatrix}\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&0\end{pmatrix}\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}$$

ou

$$\mathbf{P}=\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}$$

e

$$\mathbf{Q}=\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&0\end{pmatrix}\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix}=\begin{pmatrix}0&0&0&1\\0&1&0&0\\1&0&0&0\\0&0&1&0\end{pmatrix}$$

Logo:

$$\hat{\mathbf{A}}=\mathbf{P}\,\mathbf{A}\,\mathbf{Q}=\begin{pmatrix}6&5&-2&2\\3&2&5&10\\0&2&1&12\\0&1&0&10\end{pmatrix},\hat{\mathbf{b}}=\mathbf{P}\,\mathbf{b}=\begin{pmatrix}22\\43\\20\\12\end{pmatrix}\text{ e }\hat{\mathbf{x}}=\begin{pmatrix}x_3\\x_2\\x_4\\x_1\end{pmatrix}\Rightarrow\mathbf{x}=\mathbf{Q}\,\hat{\mathbf{x}}.$$

(a) Método de eliminação por triangularização.

$$\begin{pmatrix}6&5&-2&2&22\\3&2&5&10&43\\0&2&1&12&20\\0&1&0&10&12\end{pmatrix}$$

Normalização dos elementos da primeira linha, dividindo-os pelo primeiro elemento da mesma:

$$\begin{pmatrix}1&5/6&-1/3&1/3&11/3\\3&2&5&10&43\\0&2&1&12&20\\0&1&0&10&12\end{pmatrix}$$

Eliminação dos elementos da primeira coluna na segunda linha:

$$\begin{pmatrix}1&5/6&-1/3&1/3&11/3\\0&-1/2&6&9&32\\0&2&1&12&20\\0&1&0&10&12\end{pmatrix}$$

Normalização dos elementos da segunda linha, dividindo-os pelo segundo elemento da mesma:

$$\begin{pmatrix}1&5/6&-1/3&1/3&11/3\\0&1&-12&-18&-64\\0&2&1&12&20\\0&1&0&10&12\end{pmatrix}$$

Eliminação dos elementos da segunda coluna da terceira e quarta linhas:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 25 & 48 & 148 \\ 0 & 0 & 12 & 28 & 76 \end{pmatrix}$$

Normalização dos elementos da terceira linha, dividindo-os pelo terceiro elemento da mesma:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 12 & 28 & 76 \end{pmatrix}$$

Eliminação dos elementos da terceira coluna na quarta linha:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 0 & 124/25 & 124/25 \end{pmatrix}$$

Normalização do elemento da quarta linha, dividindo-o pelo quarto elemento da mesma:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \text{ em que o novo vetor de incógnitas é: } \hat{\mathbf{x}} = \begin{pmatrix} x_3 \\ x_2 \\ x_4 \\ x_1 \end{pmatrix}$$

Esta última forma da matriz aumentada traduz o sistema linear (triangular superior):

$$\begin{cases} x_3 + \dfrac{5}{6}x_2 - \dfrac{1}{3}x_4 + \dfrac{1}{3}x_1 = \dfrac{11}{3} \\ x_2 - 12x_4 - 18x_1 = -64 \\ x_4 + \dfrac{48}{25}x_1 = \dfrac{148}{25} \\ x_1 = 1 \end{cases}$$

Determinação recursiva de x_4, x_2 e x_3 iniciando com x_1.

$$\begin{cases} x_1 = 1 \\ x_4 = \dfrac{148}{25} - \dfrac{48}{25}x_1 = 4 \\ x_2 = -64 + 12x_4 + 18x_1 = 2 \\ x_3 = \dfrac{11}{3} - \dfrac{5}{6}x_2 + \dfrac{1}{3}x_4 - \dfrac{1}{3}x_1 = 3 \end{cases}$$

(b) Método de eliminação por diagonalização.

$$\begin{pmatrix} 6 & 5 & -2 & 2 & 22 \\ 3 & 2 & 5 & 10 & 43 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}$$

Normalização dos elementos da primeira linha, dividindo-os pelo primeiro elemento da mesma:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 3 & 2 & 5 & 10 & 43 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}$$

Eliminação dos elementos da primeira coluna na segunda linha:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & -1/2 & 6 & 9 & 32 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}$$

Normalização dos elementos da segunda linha, dividindo-os pelo segundo elemento da mesma:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 11/3 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 2 & 1 & 12 & 20 \\ 0 & 1 & 0 & 10 & 12 \end{pmatrix}$$

Eliminação dos elementos da segunda coluna da primeira, terceira e quarta linhas:

$$\begin{pmatrix} 1 & 0 & 29/3 & 46/3 & 57 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 25 & 48 & 148 \\ 0 & 0 & 12 & 28 & 76 \end{pmatrix}$$

Normalização dos elementos da terceira linha, dividindo-os pelo terceiro elemento da mesma:

$$\begin{pmatrix} 1 & 0 & 29/3 & 46/3 & 57 \\ 0 & 1 & -12 & -18 & -64 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 12 & 28 & 76 \end{pmatrix}$$

Eliminação dos elementos da terceira coluna da primeira, segunda e quarta linhas:

$$\begin{pmatrix} 1 & 0 & 0 & -242/75 & -17/75 \\ 0 & 1 & 0 & 126/25 & 176/25 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 0 & 124/25 & 124/25 \end{pmatrix}$$

Normalização do elemento da quarta linha, dividindo-o pelo quarto elemento da mesma:

$$\begin{pmatrix} 1 & 0 & 0 & -242/75 & -17/75 \\ 0 & 1 & 0 & 126/25 & 176/25 \\ 0 & 0 & 1 & 48/25 & 148/25 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Eliminação dos elementos da quarta coluna da primeira, segunda e terceira linhas:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 3 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} x_3 \\ x_2 \\ x_4 \\ x_1 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 1 \end{pmatrix}$$

(c) Método de diagonalização de Gauss para obtenção da matriz inversa. Neste caso, deseja-se determinar a matriz $\mathbf{B} = \mathbf{A}^{-1}$ tal que $\mathbf{A}\,\mathbf{B} = \mathbf{B}\,\mathbf{A} = \mathbf{I}$, em que $\mathbf{I}$ é a matriz identidade de mesma dimensão da matriz $\mathbf{A}$. Sendo, neste caso, a *matriz aumentada* é: $\mathbf{A}_{aumentada} = [\mathbf{A}\,|\,\mathbf{I}]$. Se a matriz $\mathbf{A}$ foi submetida a um procedimento de pivotamento

$$\hat{\mathbf{A}} = \mathbf{P}\,\mathbf{A}\,\mathbf{Q} \Rightarrow \hat{\mathbf{A}}_{aumentada} = [\hat{\mathbf{A}}\,|\,\mathbf{P}] \text{ e } \mathbf{A^{-1}} = \mathbf{Q}\,\hat{\mathbf{A}}^{-1}.$$

Buscando a inversa de: $\mathbf{A} = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 2 & 5 & 6 & -2 \\ 12 & 2 & 0 & 1 \\ 10 & 1 & 0 & 0 \end{pmatrix}$ para esta matriz tem-se:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \text{ e } \hat{\mathbf{A}} = \mathbf{P}\,\mathbf{A}\,\mathbf{Q} = \begin{pmatrix} 6 & 5 & -2 & 2 \\ 3 & 2 & 5 & 10 \\ 0 & 2 & 1 & 12 \\ 0 & 1 & 0 & 10 \end{pmatrix}$$

Logo:

$$\hat{\mathbf{A}}_{aumentada} = \begin{pmatrix} 6 & 5 & -2 & 2 & 0 & 1 & 0 & 0 \\ 3 & 2 & 5 & 10 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 12 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 10 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Repetindo-se o procedimento do item (b) a esta nova matriz aumentada.
Normalização dos elementos da primeira linha, dividindo-os pelo primeiro elemento da mesma:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 0 & 1/6 & 0 & 0 \\ 3 & 2 & 5 & 10 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 12 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 10 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Eliminação dos elementos da primeira coluna na segunda linha:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 0 & 1/6 & 0 & 0 \\ 0 & -1/2 & 6 & 9 & 1 & -1/2 & 0 & 0 \\ 0 & 2 & 1 & 12 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 10 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Normalização dos elementos da segunda linha, dividindo-os pelo segundo elemento da mesma:

$$\begin{pmatrix} 1 & \frac{5}{6} & -\frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 1 & -12 & -18 & -2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 12 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 10 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Eliminação dos elementos da segunda coluna da primeira, terceira e quarta linhas:

$$\begin{pmatrix} 1 & 5/6 & -1/3 & 1/3 & 0 & 1/6 & 0 & 0 \\ 0 & 1 & -12 & -18 & -2 & 1 & 0 & 0 \\ 0 & 0 & 25 & 48 & 4 & -2 & 1 & 0 \\ 0 & 0 & 0 & 12 & 28 & 2 & -1 & 0 & 1 \end{pmatrix}$$

Normalização dos elementos da terceira linha, dividindo-os pelo terceiro elemento da mesma:

$$\begin{pmatrix} 1 & 0 & 29/3 & 46/3 & 0 & 5/3 & -2/3 & 0 \\ 0 & 1 & -12 & -18 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 48/25 & 4/25 & -2/25 & 1/25 & 0 \\ 0 & 0 & 12 & 28 & 2 & -1 & 0 & 1 \end{pmatrix}$$

Eliminação dos elementos da terceira coluna da primeira, segunda e quarta linhas:

$$\begin{pmatrix} 1 & 0 & 0 & -242/75 & 3/25 & 8/75 & -29/75 & 0 \\ 0 & 1 & 0 & 126/25 & -2/25 & 1/25 & 12/25 & 0 \\ 0 & 0 & 1 & 48/25 & 4/25 & -2/25 & 1/25 & 0 \\ 0 & 0 & 0 & 124/25 & 2/25 & -1/25 & -12/25 & 1 \end{pmatrix}$$

Normalização do elemento da quarta linha, dividindo-o pelo quarto elemento da mesma:

$$\begin{pmatrix} 1 & 0 & 0 & -242/75 & 3/25 & 8/75 & -29/75 & 0 \\ 0 & 1 & 0 & 126/25 & -2/25 & 1/25 & 12/25 & 0 \\ 0 & 0 & 1 & 48/25 & 4/25 & -2/25 & 1/25 & 0 \\ 0 & 0 & 0 & 1 & 1/62 & -1/124 & -3/31 & 25/124 \end{pmatrix}$$

Eliminação dos elementos da quarta coluna da primeira, segunda e terceira linhas:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 16/93 & 5/62 & -65/93 & 121/186 \\ 0 & 1 & 0 & 0 & -5/31 & 5/62 & 30/31 & -63/62 \\ 0 & 0 & 1 & 0 & 4/31 & -2/31 & 7/31 & -12/31 \\ 0 & 0 & 0 & 1 & 1/62 & -1/124 & -3/31 & 25/124 \end{pmatrix}$$

Montando a matriz quadrada com as 4 últimas colunas da matriz acima:

$$\mathbf{M} = \begin{pmatrix} 16/93 & 5/62 & -65/93 & 121/186 \\ -5/31 & 5/62 & 30/31 & -63/62 \\ 4/31 & -2/31 & 7/31 & -12/31 \\ 1/62 & -1/124 & -3/31 & 25/124 \end{pmatrix}$$

Pré-multiplicando esta matriz por: $\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ resulta:

$$\mathbf{H} = \mathbf{Q}\,\mathbf{M} = \begin{pmatrix} 1/62 & -1/124 & -3/31 & 25/124 \\ -5/31 & 5/62 & 30/31 & -63/62 \\ 16/93 & 5/62 & -65/93 & 121/186 \\ 4/31 & -2/31 & 7/31 & -12/31 \end{pmatrix},$$

Verificando-se: $\mathbf{H}\,\mathbf{A} = \mathbf{A}\,\mathbf{H} = \mathbf{I} \Rightarrow \mathbf{H} = \mathbf{A}^{-1}$.

■

5.4 Método de Fatoração LU

O processo de fatoração LU consiste na decomposição da matriz $\mathbf{A}$ em uma matriz triangular inferior, $\mathbf{L}$, e outra triangular superior, $\mathbf{U}$, com elementos unitários na diagonal principal da matriz $\mathbf{L}$ (método de Doolittle[3]) ou da matriz $\mathbf{U}$ (método de Crout[4]):

$$\mathbf{A} = \mathbf{L}\,\mathbf{U} \Rightarrow \mathbf{A}\,\mathbf{x} = \mathbf{L}\,(\mathbf{U}\,\mathbf{x}) = \mathbf{b}$$

[3] Myrick Hascall Doolittle (1830–1911).
[4] Prescott Crout (1907–1984).

$$\text{Resultando em: } \begin{cases} \mathbf{L}\,\mathbf{y} = \mathbf{b} \\ \mathbf{U}\,\mathbf{x} = \mathbf{y} \end{cases}$$

Identificando:

$$\mathbf{L} = \begin{pmatrix} \lambda_{1,1} & 0 & 0 & \cdots & 0 \\ \lambda_{2,1} & \lambda_{2,2} & 0 & \cdots & 0 \\ \lambda_{3,1} & \lambda_{3,2} & \lambda_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & :::: & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \lambda_{n,3} & \cdots & \lambda_{n,n} \end{pmatrix} \text{ e } \mathbf{U} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \cdots & \mu_{1,n} \\ 0 & \mu_{2,2} & \mu_{2,3} & \cdots & \mu_{2,n} \\ 0 & 0 & \mu_{3,3} & \cdots & \mu_{3,n} \\ \vdots & \vdots & \vdots & :::: & \vdots \\ 0 & 0 & 0 & \cdots & \mu_{n,n} \end{pmatrix}$$

Esses dois sistemas lineares podem ser resolvidos recursivamente, por substituições direta e reversa, respetivamente:

$$\begin{cases} y_1 = \dfrac{b_1}{\lambda_{1,1}} \\ y_2 = \dfrac{b_2 - \lambda_{2,1} y_1}{\lambda_{2,2}} \\ y_3 = \dfrac{b_3 - \lambda_{3,1} y_1 - \lambda_{3,2} y_2}{\lambda_{3,3}} \\ :::: \\ y_k = \dfrac{b_k - \displaystyle\sum_{j=1}^{k-1} \lambda_{k,j} y_j}{\lambda_{k,k}} \\ :::: \\ y_n = \dfrac{b_n - \displaystyle\sum_{j=1}^{n-1} \lambda_{n,j} y_j}{\lambda_{n,n}} \end{cases} \quad \text{e} \quad \begin{cases} x_n = \dfrac{y_n}{\mu_{n,n}} \\ x_{n-1} = \dfrac{y_{n-1} - \mu_{n-1,n} x_n}{\mu_{n-1,n-1}} \\ x_{n-2} = \dfrac{y_{n-2} - \mu_{n-2,n} x_n - \mu_{n-2,n-1} x_{n-1}}{\mu_{n-2,n-2}} \\ :::: \\ x_k = \dfrac{y_k - \displaystyle\sum_{j=k+1}^{n} \mu_{k,j} x_j}{\mu_{k,k}} \\ :::: \\ x_1 = \dfrac{y_1 - \displaystyle\sum_{j=2}^{n} \mu_{1,j} x_j}{\mu_{1,1}} \end{cases}$$

No método de Doolittle tem-se $\lambda_{i,i} = 1$ e no método de Crout tem-se $\mu_{i,i} = 1$ para $i = 1, 2 \cdots n$ e em, ambos os casos, os demais elementos de **L** e **U** são determinados por:

Método de Doolittle

$$\begin{cases} \mu_{1,i} = a_{1,i} \text{ para } i = 1, 2 \cdots n \\ \lambda_{i,1} = \dfrac{a_{i,1}}{\mu_{1,1}} \text{ para } i = 2, 3, \cdots n \\ \text{para } k = 2, 3, \cdots, n: \\ \mu_{k,i} = a_{k,i} - \displaystyle\sum_{j=1}^{k-1} \lambda_{k,j} \mu_{j,i} \text{ para } i = k \cdots n \\ \lambda_{i,k} = \dfrac{a_{i,k} - \displaystyle\sum_{j=1}^{k-1} \lambda_{i,j} \mu_{j,k}}{\mu_{k,k}} \text{ para } i = k+1 \cdots n \end{cases}$$

Método de Crout

$$\begin{cases} \lambda_{i,1} = a_{i,1} \text{ para } i = 1,, 2 \cdots n \\ \mu_{1,i} = \dfrac{a_{1,i}}{\lambda_{1,1}} \text{ para } i = 2, 3 \cdots n \\ \text{para } k = 2, 3, \cdots, n: \\ \lambda_{i,k} = a_{i,k} - \displaystyle\sum_{j=1}^{k-1} \lambda_{i,j} \mu_{j,k} \text{ para } i = k \cdots n \\ \mu_{k,i} = \dfrac{a_{k,i} - \displaystyle\sum_{j=1}^{k-1} \lambda_{k,j} \mu_{j,i}}{\lambda_{k,k}} \text{ para } i = k+1 \cdots n \end{cases}$$

As principais vantagens da fatoração em relação à eliminação Gaussiana é a redução do número de operações de $\frac{2}{3}n^3 + \mathbb{O}(n^2)$ para $\frac{1}{3}n^3 + \mathbb{O}(n^2)$, e a manutenção das operações básicas na matriz fatorada (matriz **L**, na fatoração **LU**), que pode ser aplicada para diferentes vetores **b**. O pivotamento da matriz **A** também deve ser verificado para a aplicação da fatoração **LU**).

■ **Exemplo 5.3** Ilustração do método de fatoração LU pelo algoritmo de Doolitle, aplicado ao mesmo sistema do Exemplo 5.2:

$$\mathbf{A} = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 2 & 5 & 6 & -2 \\ 12 & 2 & 0 & 1 \\ 10 & 1 & 0 & 0 \end{pmatrix} \text{ e } \mathbf{b} = \begin{pmatrix} 43 \\ 22 \\ 20 \\ 12 \end{pmatrix}.$$

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda_{2,1} & 1 & 0 & 0 \\ \lambda_{3,1} & \lambda_{3,2} & 1 & 0 \\ \lambda_{4,1} & \lambda_{4,2} & \lambda_{4,3} & 1 \end{pmatrix} \text{ e } \mathbf{U} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \mu_{1,4} \\ 0 & \mu_{2,2} & \mu_{2,3} & \mu_{2,4} \\ 0 & 0 & \mu_{3,3} & \mu_{3,4} \\ 0 & 0 & 0 & \mu_{4,4} \end{pmatrix}$$

$$\mathbf{L}\cdot\mathbf{U} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \mu_{1,4} \\ \mu_{1,1}\lambda_{2,1} & \mu_{2,2}+\mu_{1,2}\lambda_{2,1} & \mu_{2,3}+\mu_{1,3}\lambda_{2,1} & \mu_{2,4}+\mu_{1,4}\lambda_{2,1} \\ \mu_{1,1}\lambda_{3,1} & \mu_{1,2}\lambda_{3,1}+\mu_{2,2}\lambda_{3,2} & \mu_{3,3}+\mu_{1,3}\lambda_{3,1}+\mu_{2,3}\lambda_{3,2} & \mu_{3,4}+\mu_{1,4}\lambda_{3,1}+\mu_{2,4}\lambda_{3,2} \\ \mu_{1,1}\lambda_{4,1} & \mu_{1,2}\lambda_{4,1}+\mu_{2,2}\lambda_{4,2} & \mu_{1,3}\lambda_{4,1}+\mu_{2,3}\lambda_{4,2}+\mu_{3,3}\lambda_{4,3} & \mu_{1,4}\lambda_{4,1}+\mu_{2,4}\lambda_{4,2}+\mu_{3,4}\lambda_{4,3}+\mu_{4,4} \end{pmatrix}$$

$$\begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \mu_{1,4} \end{pmatrix} = \begin{pmatrix} 10 & 2 & 3 & 5 \end{pmatrix}; \begin{pmatrix} \lambda_{2,1} \\ \lambda_{3,1} \\ \lambda_{4,1} \end{pmatrix} = \frac{1}{10}\begin{pmatrix} 2 \\ 12 \\ 10 \end{pmatrix} = \begin{pmatrix} 1/5 \\ 6/5 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} \mu_{2,2}+\mu_{1,2}\lambda_{2,1} & \mu_{2,3}+\mu_{1,3}\lambda_{2,1} & \mu_{2,4}+\mu_{1,4}\lambda_{2,1} \end{pmatrix} = \begin{pmatrix} 5 & 6 & -2 \end{pmatrix}$$

$$\begin{pmatrix} \mu_{2,2} & \mu_{2,3} & \mu_{2,4} \end{pmatrix} = \begin{pmatrix} 23/5 & 27/5 & -3 \end{pmatrix}$$

$$\begin{pmatrix} \mu_{1,2}\lambda_{3,1}+\mu_{2,2}\lambda_{3,2} \\ \mu_{1,2}\lambda_{4,1}+\mu_{2,2}\lambda_{4,2} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \lambda_{3,2} \\ \lambda_{4,2} \end{pmatrix} = -\begin{pmatrix} 2/23 \\ 5/23 \end{pmatrix}$$

$$\begin{pmatrix} \mu_{3,3}+\mu_{1,3}\lambda_{3,1}+\mu_{2,3}\lambda_{3,2} & \mu_{3,4}+\mu_{1,4}\lambda_{3,1}+\mu_{2,4}\lambda_{3,2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \mu_{2,3} & \mu_{3,4} \end{pmatrix} = -\begin{pmatrix} 72/23 & 121/23 \end{pmatrix}$$

$$\mu_{1,3}\lambda_{4,1}+\mu_{2,3}\lambda_{4,2}+\mu_{3,3}\lambda_{4,3} = 0 \Rightarrow \lambda_{4,3} = 7/12$$

$$\mu_{1,4}\lambda_{4,1}+\mu_{2,4}\lambda_{4,2}+\mu_{3,4}\lambda_{4,3}+\mu_{4,4} = 0 \Rightarrow \mu_{4,4} = -31/12$$

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/5 & 1 & 0 & 0 \\ 6/5 & -2/23 & 1 & 0 \\ 1 & -5/23 & 7/12 & 1 \end{pmatrix} \text{ e } \mathbf{U} = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 0 & 23/5 & 27/5 & -3 \\ 0 & 0 & -72/23 & -121/23 \\ 0 & 0 & 0 & -31/12 \end{pmatrix}$$

Para verificar se a fatoração está correta, o produto $\mathbf{L}\,\mathbf{U}$ deve ser igual à matriz $\mathbf{A}$.

$$\mathbf{L}\,\mathbf{y} = \mathbf{b} \Rightarrow \begin{cases} y_1 = 43 \\ y_2 + \dfrac{y_1}{5} = 23 \Rightarrow y_2 = \dfrac{67}{5} \\ y_3 + \dfrac{6y_1}{5} - \dfrac{2y_2}{23} = 20 \Rightarrow y_3 = -\dfrac{700}{23} \\ y_4 - \dfrac{5y_1}{23} + \dfrac{7y_2}{12} + y_3 = 12 \Rightarrow y_4 = -\dfrac{31}{3} \end{cases}$$

$$\mathbf{U}\,\mathbf{x} = \mathbf{y} \Rightarrow \begin{cases} -\dfrac{31x_4}{12} = -\dfrac{31}{3} \Rightarrow x_4 = 4 \\ -\dfrac{72x_3}{23} - \dfrac{121x_4}{23} = -\dfrac{700}{23} \Rightarrow x_3 = 3 \\ \dfrac{23x_2}{5} + \dfrac{27x_3}{5} - 3x_4 = \dfrac{67}{5} \Rightarrow x_2 = 2 \\ 10x_1 + 2x_2 + 3x_3 + 5x_4 = 43 \Rightarrow x_1 = 1 \end{cases}$$

Caso fosse necessário resolver o sistema para um valor diferente do vetor **b**, bastaria repetir os dois últimos passos do procedimento, pois a matriz **A** já está fatorada. Do mesmo modo, para obter a inversa da matriz **A**, basta repetir estes dois últimos passos para os quatro vetores coluna da matriz identidade. ∎

No caso particular de a matriz **A** ser simétrica e positiva definida, a fatoração **LU** pode ser feita pelo método de Cholesky[5], que consiste em:

$$\mathbf{A} = \mathbf{L}\,\mathbf{L}^{\mathbf{T}}.$$

Neste caso, os elementos da matriz triangular inferior **L** são determinados a partir de:

$$\begin{cases} \lambda_{1,1} = \sqrt{a_{1,1}} \\ \lambda_{i,1} = a_{i,1} \text{ para } i = 2, \cdots, n \\ \text{Para } k = 2, \cdots, n: \\ \lambda_{k,k} = \sqrt{a_{1,1} - \displaystyle\sum_{m=1}^{k-1} (\lambda_{k,m})^2} \\ \lambda_{i,k} = \dfrac{a_{i,k} - \displaystyle\sum_{m=1}^{k-1} (\lambda_{i,m}\lambda_{k,m})}{\lambda_{k,k}} \text{ para } i = k+1, \cdots, n \end{cases}$$

$$\text{Para a matriz } \mathbf{A} = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 7 \end{pmatrix} \Rightarrow \mathbf{L} = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 2\sqrt{3}/3 & \sqrt{42}/3 & 0 \\ \sqrt{3} & 2\sqrt{42}/7 & 2\sqrt{7}/7 \end{pmatrix}.$$

Caso a matriz **A** não for simétrica e positiva definida o sistema original é modificado pela multiplicação de ambos os membros por $\mathbf{A}^{\mathbf{T}}$ resultando em $\mathbf{M}\,\mathbf{x} = \mathbf{c}$, em que $\mathbf{M} = \mathbf{A}^{\mathbf{T}}\,\mathbf{A}$ e $\mathbf{c} = \mathbf{A}^{\mathbf{T}}\,\mathbf{b}$, sendo **M** uma matriz simétrica e positiva definida, possibilitando a aplicação do método de Cholesky ao sistema modificado. Aplicando este procedimento para o mesmo sistema do Exemplo 5.2, tem-se:

$$\mathbf{M} = \mathbf{A}^{\mathbf{T}}\,\mathbf{A} = \begin{pmatrix} 348 & 64 & 42 & 58 \\ 64 & 34 & 36 & 2 \\ 42 & 36 & 45 & 3 \\ 58 & 2 & 3 & 30 \end{pmatrix} \text{ e } \mathbf{c} = \mathbf{A}^{\mathbf{T}}\,\mathbf{b} = \begin{pmatrix} 834 \\ 248 \\ 261 \\ 191 \end{pmatrix}.$$

$$\mathbf{L} = \begin{pmatrix} 18,65476 & 0 & 0 & 0 \\ 3,43076 & 4,71486 & 0 & 0 \\ 2,25144 & 5,99718 & 1,99119 & 0 \\ 3,10913 & -1,83816 & 3,52743 & 2,12409 \end{pmatrix}$$

$$\mathbf{L}\,\mathbf{y} = \mathbf{c} \Rightarrow \mathbf{y} = \begin{pmatrix} 44,70709 \\ 20,06862 \\ 20,08329 \\ 8,49634 \end{pmatrix} \text{ e } \mathbf{L}^{\mathbf{T}}\,\mathbf{x} = \mathbf{y} \Rightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

5.5 Método de Thomas para Matrizes Tridiagonais

Método de Thomas[6] ou TDMA (*Tri-Diagonal Matrix Algorithm*): Um caso particular, muito comum, de sistemas lineares, é o sistema tri-diagonal. Este tipo de sistema surge em problemas de

[5] André-Louis Cholesky (1875–1918).
[6] Llewellyn Hilleth Thomas (1903-1992).

Engenharia Química envolvendo, por exemplo, a modelagem estacionária de sistemas em estágios e a discretização de equações diferenciais por diferenças finitas. Sistemas tri-diagonais podem ser expressos genericamente na forma indicial:

$$\begin{cases} b_0 x_0 + c_0 x_1 = d_0 \\ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i \text{ para } i = 1, 2, \cdots, n-1 \\ a_n x_{n-1} + b_n x_n = d_n \end{cases}$$

O método de Thomas aplicado à resolução deste tipo de sistema é uma versão da decomposição **LU** adaptada à estrutura esparsa da matriz do sistema, assim é proposta a seguinte recorrência como solução:

$$x_i = \gamma_i - \frac{c_i}{\beta_i} x_{i+1}.$$

Como: $b_0 x_0 + c_0 x_1 = d_0 \Rightarrow x_0 = \dfrac{d_0}{b_0} - \dfrac{c_0}{b_0} x_1 = \gamma_0 - \dfrac{c_0}{\beta_0} x_1$, logo $\beta_0 = b_0$ e $\gamma_0 = \dfrac{d_0}{b_0}$.
Substituindo a recorrência em:

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i \Rightarrow a_i \left[\gamma_{i-1} - \frac{c_{i-1}}{\beta_{i-1}} x_i \right] + b_i x_i = d_i - c_i x_{i+1}$$

e agrupando os coeficientes da equação, obtém-se:

$$\left[b_i - \frac{a_i c_{i-1}}{\beta_{i-1}} \right] x_i = (d_i - a_i \gamma_{i-1}) - c_i x_{i+1}$$

Permitindo identificar: $\beta_i = b_i - \dfrac{a_i c_{i-1}}{\beta_{i-1}}$ e $\gamma_i = \dfrac{d_i - a_i \gamma_{i-1}}{\beta_i}$ para $i = 1, 2, \cdots, n$.

Dessa forma, os coeficientes β_i e γ_i são determinados recursivamente na forma:

$$\text{Para } i = 1, 2, \cdots, n \begin{cases} \beta_i = b_i - \dfrac{a_i c_{i-1}}{\beta_{i-1}} \text{ com } \beta_0 = b_0 \\ \gamma_i = \dfrac{d_i - a_i \gamma_{i-1}}{\beta_i} \text{ com } \gamma_0 = \dfrac{d_0}{b_0} \end{cases}$$

e as variáveis x_i são determinadas pela forma recursiva:

$$x_i = \gamma_i - \frac{c_i}{\beta_i} x_{i+1} \text{ para } i = n-1, n-2, \cdots, 0, \text{ com } x_n = \gamma_n.$$

Na modelagem estacionária de sistemas em estágios com reciclos origina-se uma forma *perturbada* de sistemas tri-diagonais descrita por:

$$\begin{cases} a_0 x_n + b_0 x_0 + c_0 x_1 = d_0 \\ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i \text{ para } i = 1, 2, \cdots, n-1 \\ a_n x_{n-1} + b_n x_n + c_n x_0 = d_n \end{cases}$$

O método de resolução desta forma modificada de sistemas lineares tri-diagonais é descrita em https://www.cfd-online.com/Wiki/Tridiagonal_matrix_algorithm_-_TDMA_(Thomas_algorithm) em que é aplicada a fórmula de Sherman e Morrison (1950).

O método consiste em transformar a matriz original em uma nova matriz tri-diagonal não perturbada através da subtração:

$$\hat{\mathbf{A}} = \mathbf{A} - \mathbf{u}\,\mathbf{v}^{\mathbf{T}} \text{ em que } \mathbf{u}^{\mathbf{T}} = \begin{pmatrix} -b_1 & 0 & 0 & \cdots & c_n \end{pmatrix}, \mathbf{v}^{\mathbf{T}} = \begin{pmatrix} 1 & 0 & 0 & \cdots & -\dfrac{a_1}{b_1} \end{pmatrix}$$

$$\text{e } \mathbf{A} = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 & a_1 \\ a_2 & b_2 & c_2 & \cdots & 0 & 0 \\ 0 & a_3 & b_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} & c_{n-1} \\ c_n & 0 & 0 & \cdots & a_n & b_n \end{pmatrix}$$

$$\text{Assim: } \mathbf{u}\mathbf{v}^{\mathbf{T}} = \begin{pmatrix} -b_1 & 0 & 0 & \cdots & a_1 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ c_n & 0 & 0 & \cdots & -\dfrac{c_n a_1}{b_1} \end{pmatrix} \text{ e } \hat{\mathbf{A}} = \begin{pmatrix} 2b_1 & c_1 & 0 & \cdots & 0 & 0 \\ a_2 & b_2 & c_2 & \cdots & 0 & 0 \\ 0 & a_3 & b_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & \cdots & a_n & b_n + \dfrac{c_n a_1}{b_1} \end{pmatrix}$$

Resolvendo os dois sistemas tri-diagonais: $\hat{\mathbf{A}}\,\mathbf{y} = \mathbf{d}$ e $\hat{\mathbf{A}}\,\mathbf{z} = \mathbf{u}$, a solução do sistema original será:

$$\mathbf{x} = \mathbf{y} - \left(\frac{\mathbf{v}^{\mathbf{T}}\,\mathbf{y}}{1 + \mathbf{v}^{\mathbf{T}}\,\mathbf{z}} \right) \mathbf{z}.$$

Sistemas lineares penta-diagonais aparecem no processo de discretização de equações diferenciais pelo método de diferenças finitas de quarta ordem , tais sistemas podem ser expressos genericamente na forma indicial:

$$\begin{cases} c_0 x_0 + d_0 x_1 + e_0 x_2 = y_0 \\ b_1 x_0 + c_1 x_1 + d_1 x_2 + e_1 x_3 = y_1 \\ a_i x_{i-2} + b_i x_{i-1} + c_i x_i + d_i x_{i+1} + e_i x_{i+2} = y_i \text{ para } i = 2, \cdots, n-2 \\ a_{n-1} x_{n-3} + b_{n-1} x_{n-2} + c_{n-1} x_{n-1} + d_{n-1} x_n = y_{n-1} \\ a_n x_{n-2} + b_n x_{n-1} + c_n x_n = y_n \end{cases}$$

Um procedimento semelhante ao método de Thomas é proposto para a resolução desse tipo de sistema, assim é proposta a seguinte forma triangular superior bi-diagonal como solução: $x_i + \beta_i x_{i+1} + \delta_i x_{i+2} = z_i$ para $i = 0, 1, \cdots (n-2), x_{n-1} + \beta_{n-1} x_n = z_{n-1}$ e $x_n = z_n$. Da primeira equação do sistema conclui-se que: $\beta_0 = \dfrac{d_0}{c_0}$, $\delta_0 = \dfrac{e_0}{c_0}$ e $z_0 = \dfrac{y_0}{c_0}$. Substituindo na segunda equação $x_0 = z_0 - \beta_0 x_1 - \delta_0 x_2$, resulta:

$$\beta_1 = \frac{d_1 - b_1\,\delta_0}{c_1 - b_1\,\beta_0},\ \delta_1 = \frac{e_1}{c_1 - b_1\,\beta_0} \text{ e } z_1 = \frac{y_1 - b_1 z_0}{c_1 - b_1\,\beta_0}.$$

Substituindo na i-ésima equação $x_{i-1} = z_{i-1} - \beta_{i-1} x_i - \delta_{i-1} x_{i+1}$, e

$$x_{i-2} = z_{i-2} - \beta_{i-2} x_{i-1} - \delta_{i-2} x_i = z_{i-2} - \beta_{i-2} \left(z_{i-1} - \beta_{i-1} x_i - \delta_{i-1} x_{i+1} \right) - \delta_{i-2} x_i,$$

$$\text{resulta: } \beta_i = \frac{d_i - b_i\,\delta_{i-1} + a_i\,\beta_{i-2}\,\delta_{i-1}}{c_i - b_i\,\beta_{i-1} - a_i\,\delta_{i-2} + a_i\,\beta_{i-2}\,\beta_{i-1}},\ \delta_i = \frac{e_i}{c_i - b_i\,\beta_{i-1} - a_i\,\delta_{i-2} + a_i\,\beta_{i-2}\,\beta_{i-1}}$$

$$\text{e } z_i = \frac{y_i - (a_i z_{i-2} + b_i z_{i-1} - a_i\,\beta_{i-2} z_{i-1})}{c_i - b_i\,\beta_{i-1} - a_i\,\delta_{i-2} + a_i\,\beta_{i-2}\,\beta_{i-1}} \text{ para } i = 2, \cdots, n, \text{ com } e_{n-1} = e_n = d_n = 0.$$

Após a determinação de β_i, δ_i e z_i para $i = 0, 1, \cdots, n$ aplica-se a forma recursiva:

$$\begin{cases} x_n = z_n \\ x_{n-1} = z_{n-1} - \beta_{n-1} x_n \\ x_i = z_i - \beta_i x_{i+1} - \delta_i x_{i+2} \text{ para } i = (n-2), \cdots, 1, 0. \end{cases}$$

■ **Exemplo 5.4** Polinômio Interpolador *spline*. Uma aplicação importante de sistemas lineares tri-diagonais são as funções interpoladoras de terceiro grau do tipo *spline*. Tais funções polinomiais são seccionalmente contínuas e passam por $(n+1)$ pontos nodais (x_i, y_i) para $i = 0, 1, \cdots, n$, no intervalo ou elemento i da interpolação, em que $x_{i-1} \leq x \leq x_i$, os valores da coordenada y é interpolada por um polinômio de terceiro grau $S^{(i)}(x)$, tal polinômio é seccionalmente contínuo com suas duas primeiras derivadas também seccionalmente contínuas em ambas extremidades do intervalo. Estas propriedades podem ser equacionadas por:

$$\begin{cases} S^{(i)}(x_{i-1}) = y_{i-1} \text{ e } S^{(i)}(x_i) = y_i \\ S'^{(i)}(x_i) = S'^{(i+1)}(x_i) \\ S''^{(i)}(x_{i-1}) = \phi_{i-1} \text{ e } S''^{(i)}(x_i) = \phi_i \end{cases}.$$

Com tais propriedades obtém-se:

$$S^{(i)}(x) = \left(\frac{x_i - x}{h_i}\right) y_{i-1} + \left(\frac{x - x_{i-1}}{h_i}\right) y_i + \frac{(x_i - x)}{6h_i}\left[(x_i - x)^2 - h_i^2\right]\phi_{i-1} + \\ + \frac{(x - x_{i-1})}{6h_i}\left[(x - x_{i-1})^2 - h_i^2\right]\phi_i \text{ em que } h_i = x_i - x_{i-1}.$$

Aplicando a propriedade $S'^{(i)}(x_i) = S'^{(i+1)}(x_i)$ para elementos internos (não contendo as extremidades x_0 e x_n), obtém-se:

$$\frac{h_i}{6}\phi_{i-1} + \left(\frac{h_{i+1} + h_i}{3}\right)\phi_i + \frac{h_{i+1}}{6}\phi_{i+1} = \left(\frac{y_{i+1} - y_i}{h_{i+1}}\right) - \left(\frac{y_i - y_{i-1}}{h_i}\right) \; i = 1, 2, \cdots, (n-1)$$

Permitindo identificar:

$$a_i = \frac{h_i}{6},\; b_i = \left(\frac{h_{i+1} + h_i}{3}\right),\; c_i = \frac{h_{i+1}}{6} \text{ e } d_i = \left(\frac{y_{i+1} - y_i}{h_{i+1}}\right) - \left(\frac{y_i - y_{i-1}}{h_i}\right)$$

Nos dois elementos das extremidades, três hipóteses distintas são consideradas:

(a) Em x_0 e x_n a segunda derivada é nula, isto é: $\phi_0 = \phi_n = 0 \Rightarrow b_0 = b_n = 1, c_0 = d_0 = b_n = d_n = 0$.

(b) Nos dois elementos da extremidade o perfil é parabólico, isto é: $\phi_0 = \phi_1$ ou $\phi_0 - \phi_1 = 0$ e $\phi_{n-1} = \phi_n$
,$textrmou - \phi_{n-1} + \phi_n = 0 \Rightarrow b_0 = b_n = 1, c_1 = a_n = -1, d_0 = d_n = 0$.

(c) O coeficiente de x^3 em $S^{(1)}(x)$ é igual ao coeficiente de x^3 em $S^{(2)}(x)$, o mesmo valendo para $S^{(n)}(x)$ e $S^{(n-1)}(x)$. Resultando em:

$$\begin{cases} -\frac{1}{h_1}\phi_0 + \left(\frac{1}{h_1} + \frac{1}{h_2}\right)\phi_1 - \frac{1}{h_2}\phi_2 = 0 \\ -\frac{1}{h_{n-1}}\phi_{n-2} + \left(\frac{1}{h_{n-1}} + \frac{1}{h_n}\right)\phi_{n-1} - \frac{1}{h_n}\phi_n = 0 \end{cases}$$

Nesse último caso, para manter a estrutura tri-diagonal do sistema, deve-se eliminar o termo em ϕ_2 na primeira equação e o termo em ϕ_{n-2} na segunda equação. Assim:

$$\begin{cases} -\frac{h_2}{h_1}\phi_0 + \left(1 + \frac{h_2}{h_1}\right)\phi_1 - \phi_2 = 0 \\ a_1\,\phi_0 + b_1\,\phi_1 + c_1\,\phi_2 = d_1 \end{cases} \Rightarrow b_0\,\phi_0 + c_0\,\phi_1 = d_0$$

$$b_0 = a_1 - c_1\frac{h_2}{h_1},\; c_0 = b_1 + c_1\left(1 + \frac{h_2}{h_1}\right),\; d_0 = d_1.$$

$$\begin{cases} -\phi_{n-2} + \left(1 + \dfrac{h_{n-1}}{h_n}\right)\phi_{n-1} - \dfrac{h_{n-1}}{h_n}\phi_n = 0 \\ a_{n-1}\,\phi_{n-2} + b_{n-1}\,\phi_{n-1} + c_{n-1}\,\phi_n = d_{n-1} \end{cases} \Rightarrow a_n\,\phi_{n-1} + c_n\,\phi_n = d_n$$

$$a_n = b_{n-1} + c_{n-1}\left(1 + \frac{h_{n-1}}{h_n}\right),\ c_n = c_{n-1} - a_{n-1}\frac{h_{n-1}}{h_n}\,d_n = d_{n-1}.$$

Para ilustrar o procedimento os seguintes valores numéricos são utilizados:

Ponto	0	1	2	3	4	5	6	7	8	9	10
x_i	0,000	1,040	1,901	2,801	4,122	4,884	6,021	6,846	8,101	9,017	9,975
y_i	0,000	0,184	0,092	-0,396	-0,069	0,021	0,057	-0,379	-0,076	0,016	0,005

Na Figura 5.4 representa-se a correspondente curva *spline* de terceiro grau com parábolas nos elementos extremos [Hipótese (b)].

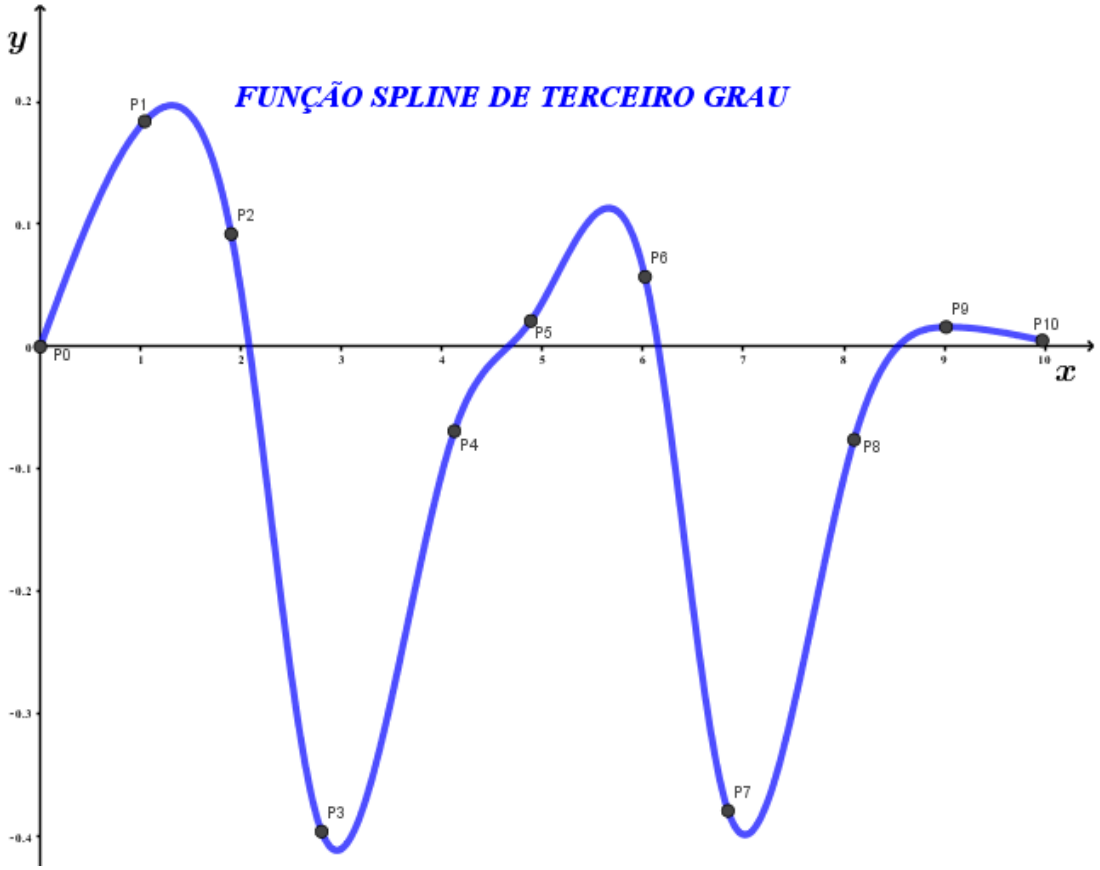


Figura 5.4: Curva *spline* de terceiro grau com parábolas nos elementos extremos.

■

5.6 Métodos Iterativos para a Resolução de Sistemas Algébricos Lineares

Técnicas iterativas são raramente utilizadas para a resolução de sistemas algébricos lineares de baixas dimensões, já que o tempo requerido para obter um mínimo de acurácia na aplicação dessas técnicas ultrapassa o tempo requerido pelas técnicas diretas do tipo da eliminação gaussiana. Contudo, para sistemas de dimensões elevadas, com grande porcentagem de elementos nulos (sistemas esparsos), tais técnicas aparecem como alternativas mais eficientes. O uso de algoritmos

de álgebra esparsa, não abordados neste texto, é outra alternativa eficiente para a resolução por métodos diretos de sistemas lineares de dimensões elevadas. Sistemas esparsos de grande porte frequentemente surgem na resolução numérica de equações diferenciais ordinárias com problemas de valor no contorno e de equações diferenciais parciais. Da mesma forma que os métodos diretos, existe uma grande variedade de métodos iterativos para resolução iterativa de sistemas algébricos lineares. A literatura desta área é muito dinâmica não sendo possível incluir a enorme gama de procedimentos em um simples texto, assim apenas os métodos mais usuais e tradicionais são aqui descritos, são estes:

- Método Iterativo de Jacobi[7].
- Método Iterativo de Gauss-Seidel[8].
- Método Iterativo tipo Sobre-Relaxação Sucessiva (Sucessive Over-Relaxation **SOR**).
- Método Iterativo Fundamentado no Método do Gradiente Conjugado.

Antes de apresentar os métodos iterativos de resolução de sistemas algébricos lineares, um resumo dos conceitos de normas de matrizes quadradas são apresentados. Tais conceitos são importantes na análise da convergência desses procedimentos iterativos. O conceito de norma de uma matriz está intimamente relacionado ao conceito de norma de vetor, assim:

$$||\mathbf{A}|| = \max \frac{||\mathbf{A}\,\mathbf{x}||}{||\mathbf{x}||}$$

Se o vetor $\mathbf{x}$ for selecionado tal que $||\mathbf{x}|| = 1$ tem-se:

$$||\mathbf{A}|| = \max ||\mathbf{A}\,\mathbf{x}||.$$

A interpretação analítica de $||\mathbf{x}|| = 1$ depende da norma do vetor, assim:

(a) Norma máxima ou infinita (l_∞) : $||\mathbf{x}||_\infty = \arg\max_{i=1}^{n} |x_i| = 1$ seriam os pontos situados sobre o hipercubo de aresta unitária;

(b) Norma absoluta (l_1) : $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i| = 1$ seriam os pontos situados sobre o hipercubo com vértices situados no eixos coordenados nas posições ± 1;

(c) Norma euclidiana (l_2) : $||\mathbf{x}||_2 = \sum_{i=1}^{n} \sqrt{(x_i)^2} = 1$ seriam os pontos situados sobre o hiperesfera de raio unitário e centro na origem.

Para melhor entender esses conceitos uma matriz (2×2) é considerada:

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$$

Vetores unitários em relação à l_∞ :

$$\mathbf{v_1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{v_2} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \mathbf{v_3} = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \text{ e } \mathbf{v_4} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Assim:

$$\mathbf{A}\,\mathbf{v_1} = \begin{pmatrix} a_{1,1}+a_{1,2} \\ a_{2,1}+a_{2,2} \end{pmatrix}, \mathbf{A}\,\mathbf{v_2} = \begin{pmatrix} -a_{1,1}+a_{1,2} \\ -a_{2,1}+a_{2,2} \end{pmatrix}, \mathbf{A}\,\mathbf{v_3} = \begin{pmatrix} -a_{1,1}-a_{1,2} \\ -a_{2,1}-a_{2,2} \end{pmatrix} \text{ e } \mathbf{A}\,\mathbf{v_4} = \begin{pmatrix} a_{1,1}-a_{1,2} \\ a_{2,1}-a_{2,2} \end{pmatrix}$$

Permitindo inferir: $||\mathbf{A}||_\infty = \max\left(|a_{1,1}|+|a_{1,2}|, |a_{2,1}|+|a_{2,2}|\right)$, isto é, $||\mathbf{A}||_\infty$ é igual ao máximo da soma dos módulos dos elementos da mesma linha. Para ilustrar esse conceito a matriz $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix}$ é considerada, assim: $||\mathbf{A}||_\infty = \max(1+2, 1+4) = 5$.

[7] Carl Gustav Jacob Jacobi (1804-1851).

[8] Philipp Ludwig von Seidel (1821-1896).

Vetores unitários em relação à l_1 :

$$\mathbf{v_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{v_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mathbf{v_3} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \text{ e } \mathbf{v_4} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

Assim:

$$\mathbf{A\,v_1} = \begin{pmatrix} a_{1,1} \\ a_{2,1} \end{pmatrix}, \mathbf{A\,v_2} = \begin{pmatrix} a_{1,2} \\ a_{2,2} \end{pmatrix}, \mathbf{A\,v_3} = \begin{pmatrix} -a_{1,1} \\ -a_{2,1} \end{pmatrix} \text{ e } \mathbf{A\,v_4} = \begin{pmatrix} -a_{1,2} \\ -a_{2,2} \end{pmatrix}$$

Permitindo inferir: $||\mathbf{A}||_\mathbf{1} = \max\left(|a_{1,1}| + |a_{2,1}|, |a_{1,2}| + |a_{2,2}|\right)$, isto é, $||\mathbf{A}||_\mathbf{1}$ é igual ao máximo da soma dos módulos dos elementos da mesma coluna. Para ilustrar esse conceito a mesma matriz da norma anterior é considerada, assim: $||\mathbf{A}||_\mathbf{1} = \max(1+1, 2+4) = 6$.

Vetor unitário em relação à l_2 : $\mathbf{v} = \begin{pmatrix} \cos(\theta) \\ \text{sen}(\theta) \end{pmatrix}$ então $\mathbf{A\,v} = \begin{pmatrix} a_{1,1}\cos(\theta) + a_{1,2}\,\text{sen}(\theta) \\ a_{2,1}\cos(\theta) + a_{2,2}\,\text{sen}(\theta) \end{pmatrix}$.

Permitindo inferir: $||\mathbf{A}||_\mathbf{2} = \max|\mathbf{A\,v}(\theta)|$, considerando a mesma matriz das normas anteriores: $||\mathbf{A}||_\mathbf{2} = \max\sqrt{[\cos(\theta) + 2\,\text{sen}(\theta)]^2 + [4\,\text{sen}(\theta) - \cos(\theta)]^2}$ por um procedimento numérico apropriado determina-se o valor de $\theta = 1,6801$ que corresponde a $||\mathbf{A}||_\mathbf{2} = 4,4966$. Pode-se demonstrar que este valor é a raiz quadrada do raio espectral da matriz $\mathbf{A\,A^T}$. O raio espectral de uma matriz $\mathbf{M}$ é definido por: $\rho(\mathbf{M}) = \max_{i=1}^{n}(|\lambda_i|)$ em que λ_i são os *valores característicos* da matriz (determinados por código computacional apropriado).

5.6.1 Método de Jacobi

Considerando a partição da matriz característica do sistema: $\mathbf{A} = \mathbf{D} - (\mathbf{D} - \mathbf{A})$ sendo $\mathbf{D}$ a matriz diagonal que contém os elementos da diagonal principal da matriz $\mathbf{A}$, assim:

$$\mathbf{A\,x} = \mathbf{D\,x} - (\mathbf{D} - \mathbf{A})\,\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{D^{-1}}\,(\mathbf{D} - \mathbf{A})\,\mathbf{x} + \mathbf{D^{-1}}\,\mathbf{b}.$$

Definindo: $\mathbf{M} = \mathbf{D^{-1}}\,(\mathbf{D} - \mathbf{A})$ e $\mathbf{c} = \mathbf{D^{-1}}\,\mathbf{b}$ tem-se: $\mathbf{x} = \mathbf{M\,x} + \mathbf{c}$.
Sugerindo o procedimento iterativo:

$$\mathbf{x^{(k+1)}} = \mathbf{M\,x^{(k)}} + \mathbf{c} \text{ para } k = 0, 1, \cdots,$$

que pode ser escrito na forma indicial como:

$$x_i^{(k+1)} = \frac{b_i - \sum_{j=1(\neq i)}^{n} a_{i,j}\,x_j^{(k)}}{a_{i,i}} \text{ para } i = 1, 2, \cdots, n \text{ e } k = 0, 1, \cdots.$$

5.6.2 Método de Gauss-Seidel

Considerando a partição da matriz característica do sistema: $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$ sendo $\mathbf{D}$ a matriz diagonal que contém os elementos da diagonal principal da matriz $\mathbf{A}$, $\mathbf{L}$ a matriz triangular inferior contendo os elementos sob a diagonal principal da matriz $\mathbf{A}$ e $\mathbf{U}$ a matriz triangular superior contendo os elementos sobre a diagonal principal da matriz $\mathbf{A}$, assim:

$$\mathbf{A\,x} = \mathbf{D\,x} + (\mathbf{L} + \mathbf{U})\,\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{D^{-1}}\,\mathbf{b} - \mathbf{D^{-1}}\,\mathbf{L\,x} - \mathbf{D^{-1}}\,\mathbf{U\,x}.$$

Sugerindo o procedimento iterativo: $\mathbf{x^{(k+1)} = D^{-1}\,b - D^{-1}\,L\,x^{(k+1)} - D^{-1}\,U\,x^{(k)}}$, que pode ser escrito na forma indicial como:

$$x_i^{(k+1)} = \begin{cases} \dfrac{b_1 - \sum_{j=2}^{n} a_{1,j}x_j^{(k)}}{a_{1,1}} \text{ para } i = 1 \\ \dfrac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}}{a_{i,i}} \text{ para } i = 2, \cdots, n \end{cases} \qquad k = 0, 1, \cdots$$

Neste procedimento a matriz $\mathbf{M}$ do procedimento iterativo pode ser considerada como sendo: $\mathbf{M = -(L+D)^{-1}\,U}$.

5.6.3 Método das Sobre-Relaxações Sucessivas (SOR)

Este procedimento visa acelerar a convergência do método de Gauss-Seidel pela introdução de um fator de relaxação, de acordo com o procedimento iterativo:

$$x_i^{(k+1)} = x_i^{(k)} + \omega\left(\hat{x}_i^{(k+1)} - x_i^{(k)}\right) \text{ em que } \hat{x}_i^{(k+1)} \text{ é calculado pelo método de Gauss-Seidel.}$$

O parâmetro ω é o fator de relaxação, quando $1 < \omega$ o procedimento é dito de **sobre-relaxação** (*over-relaxation*) o que acelera a convergência do método de Gauss-Seidel (caso seja convergente). Escolhendo-se $0 < \omega < 1$ o método é chamado de **sub-relaxação** e pode assegurar a convergência de procedimentos iterativos não-convergentes.

Neste procedimento a matriz $\mathbf{M}$ do procedimento iterativo pode ser considerada como sendo: $\mathbf{I - \omega\left[L+D\right)^{-1}\,U + I]}$.

A convergência destes três métodos iterativos é caracterizada pela matriz de iteração $\mathbf{M}$ sendo convergentes se, e somente se, todos os valores característicos de $\mathbf{M}$ possuírem valor absoluto menor que 1. A convergência é também assegurada se a norma de $\mathbf{M}$ for menor que 1, podendo ser computada por alguma das definições a seguir:

- Norma máxima ou infinita: $||\mathbf{M}||_\infty = \max_{i=1}^{n} \sum_{j=1}^{n} |m_{i,j}|$;
- Norma absoluta: $||\mathbf{M}||_\mathbf{1} = \max_{j=1}^{n} \sum_{i=1}^{n} |m_{i,j}|$;
- Norma euclidiana: $||\mathbf{M}||_\mathbf{2} = \max_{i=1}^{n} \sqrt{|\lambda_i|}$, sendo λ_i os valores característicos de $\mathbf{M\,M^T}$;
- Norma de Frobenius[9]: $||\mathbf{M}||_F = \sqrt{traço(\mathbf{M\,M^T})}$.

A convergência dos métodos iterativos também pode ser assegurada se a matriz $\mathbf{A}$ for diagonalmente dominante, ou seja,

$$|a_{i,i}| > \sum_{j=1(\neq i)}^{n} |a_{i,j}| \text{ para } i = 1, 2, \cdots, n.$$

Isso pode ser prontamente verificado para o método de Jacobi e usando a norma infinita, pois a expressão acima é equivalente a:

$$\sum_{j=1(\neq i)}^{n} \frac{|a_{i,j}|}{|a_{i,i}|} = \sum_{j=1}^{n} |m_{i,j}| < 1 \text{ para } i = 1, 2, \cdots, n \Rightarrow ||\mathbf{M}||_\infty < 1.$$

[9] Ferdinand Georg Frobenius (1849-1917).

5.6.4 Método Fundamentado no Método do Gradiente Conjugado

Inicialmente o problema original, $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, é transformado de forma a assegurar que a matriz característica do sistema seja positiva definida e simétrica, para isto deve-se multiplicar ambos os lados da equação pela matriz $\mathbf{A^T}$ assim:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{A^T}\,\mathbf{A}\,\mathbf{x} = \mathbf{A^T}\,\mathbf{b} \text{ definindo } \mathbf{M} = \mathbf{A^T}\,\mathbf{A} \text{ e } \mathbf{c} = \mathbf{A^T}\,\mathbf{b}, \text{ resulta: } \mathbf{M}\,\mathbf{x} = \mathbf{c}.$$

Este novo problema é então transformado em um problema de minimização da função quadrática:

$$S(\mathbf{x}) = \frac{1}{2}\mathbf{x^T}\,\mathbf{M}\,\mathbf{x} - \mathbf{c^T}\,\mathbf{x} \;\Rightarrow\; \nabla S(\mathbf{x}) = \mathbf{M}\,\mathbf{x} - \mathbf{c}$$

Considerando $\mathbf{z}$ um valor genérico da variável, tem-se:

$$S(\mathbf{z}) = \frac{1}{2}\mathbf{z^T}\,\mathbf{M}\,\mathbf{z} - \mathbf{c^T}\,\mathbf{z} \Rightarrow \nabla S(\mathbf{z}) = \mathbf{M}\,\mathbf{z} - \mathbf{c} = \mathbf{r}.$$

O processo iterativo é orientado na busca do valor de $\mathbf{z}$ que anula $\mathbf{r}$, iniciando o procedimento iterativo com um valor arbitrário de $\mathbf{z}$, $\mathbf{z^{(0)}}$ calcula-se a seguir $\mathbf{r^{(0)}} = \mathbf{M}\,\mathbf{z^{(0)}} - \mathbf{c}$. O próximo valor de $\mathbf{z}$ é buscado na direção $\mathbf{p^{(0)}}$ com um passo λ_0 a partir de $\mathbf{z^{(0)}}$, isto é:

$$\mathbf{z^{(1)}} = \mathbf{z^{(0)}} + \lambda_0\,\mathbf{p^{(0)}} \Rightarrow \mathbf{r^{(1)}} = \mathbf{r^{(0)}} + \lambda_0\,\mathbf{M}\,\mathbf{p^{(0)}}.$$

Para que $\mathbf{r^{(1)}}$ assuma o menor valor possível deve ser ortogonal a $\mathbf{p^{(0)}}$ implicando em $\lambda_0 = -\dfrac{(\mathbf{r^{(0)}})^{\mathbf{T}}\,\mathbf{p^{(0)}}}{(\mathbf{p^{(0)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(0)}}}$, o valor inicial da direção de busca, $\mathbf{p^{(0)}}$, é escolhido pelo método da descida mais íngreme (*steepest descent*): $\mathbf{p^{(0)}} = -\nabla\mathbf{S}(\mathbf{z^{(0)}}) = -\mathbf{r^{(0)}}$.

Para os passos seguintes a direção escolhida será sempre a direção *conjugada* à anterior, isto é: $(\mathbf{p^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}} = 0$ considerando esta nova direção uma *correção* do método do gradiente, $\mathbf{p^{(k)}} = -\mathbf{r^{(k)}} + \varepsilon_{\mathbf{k-1}}\mathbf{p^{(k-1)}}$ substituindo esta equação em $(\mathbf{p^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}} = 0$, resulta em $\varepsilon_{k-1} = \dfrac{(\mathbf{r^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}}}{(\mathbf{p^{(k-1)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}}}$.

A seguir calculam-se:

$$\mathbf{z^{(k+1)}} = \mathbf{z^{(k)}} + \lambda_{\mathbf{k}}\mathbf{p^{(k)}} \text{ e } \lambda_k = -\frac{(\mathbf{r^{(k)}})^{\mathbf{T}}\,\mathbf{p^{(k)}}}{(\mathbf{p^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k)}}}.$$

Resumindo:

$$\text{Início:}\begin{cases}\mathbf{z^{(0)}} \text{ (valor arbitrário)}\\ \mathbf{r^{(0)}} = \mathbf{M}\,\mathbf{z^0} - \mathbf{c}\\ \mathbf{p^{(0)}} = -\mathbf{r^{(0)}}\\ \lambda_0 = -\dfrac{(\mathbf{r^{(0)}})^{\mathbf{T}}\,\mathbf{p^{(0)}}}{(\mathbf{p^{(0)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(0)}}}\\ \mathbf{z^{(1)}} = \mathbf{z^{(0)}} + \lambda_0\mathbf{p^{(0)}}\\ \mathbf{r^{(1)}} = \mathbf{r^{(0)}} + \lambda_0\mathbf{M}\,\mathbf{p^{(0)}}.\end{cases} \qquad \text{Iteração } k = 1, 2, \cdots \begin{cases}\varepsilon_{k-1} = \dfrac{(\mathbf{r^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}}}{(\mathbf{p^{(k-1)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k-1)}}}\\ \mathbf{p^{(k)}} = -\mathbf{r^{(k)}} + \varepsilon_{\mathbf{k-1}}\mathbf{p^{(k-1)}}\\ \lambda_k = -\dfrac{(\mathbf{r^{(k)}})^{\mathbf{T}}\,\mathbf{p^{(k)}}}{(\mathbf{p^{(k)}})^{\mathbf{T}}\,\mathbf{M}\,\mathbf{p^{(k)}}}\\ \mathbf{z^{(k+1)}} = \mathbf{z^{(k)}} + \lambda_{\mathbf{k}}\mathbf{p^{(k)}}\\ \mathbf{r^{(k+1)}} = \mathbf{M}\,\mathbf{z^{(k+1)}} - \mathbf{c}.\end{cases}$$

Exemplo numérico: $\mathbf{M} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$ e $\mathbf{c} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$.

$\mathbf{z^{(0)}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathbf{r^{(0)}} = -\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$, $\mathbf{p^{(0)}} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$, $\lambda_0 = \frac{1}{2}$, $\mathbf{z^{(1)}} = \begin{pmatrix} 1{,}25 \\ 1{,}25 \end{pmatrix}$

$\mathbf{r^{(1)}} = \begin{pmatrix} -0{,}125 \\ 0{,}125 \end{pmatrix}$, $\varepsilon_0 = 0{,}0625$, $\mathbf{p^{(1)}} = \begin{pmatrix} 0{,}15625 \\ -0{,}09375 \end{pmatrix}$, $\lambda_1 = 1{,}142857$

$\mathbf{z^{(2)}} = \begin{pmatrix} 1{,}428571 \\ 1{,}142857 \end{pmatrix}$ e $\mathbf{r^{(1)}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, logo $\mathbf{z^{(2)}}$ é a solução.

5.7 Métodos para a Resolução de Sistemas Algébricos Não Lineares

Nesta seção, os métodos das substituições sucessivas, de Newton-Raphson e da secante, apresentados no Capítulo 4 para a resolução de uma única equação algébrica, são estendidos para sistemas de equações algébricas não lineares. A formulação de um problema de otimização para a resolução desses sistemas também é brevemente descrita. Finalmente o método da continuação paramétrica é introduzido para a obtenção de múltiplas soluções desses sistemas não lineares.

5.7.1 Método de Substituições Sucessivas

De modo semelhante à resolução de equações não lineares em uma variável o método de substituições sucessivas (ou pontos fixos) aplicado a sistemas algébricos não lineares: $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, consiste na transformação do problema em: $\mathbf{x} = \mathbf{g}(\mathbf{x})$, em que $\mathbf{g}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ e $\mathbf{x} \in \mathbb{R}^n$. Tal transformação sugere o procedimento iterativo:

$$\mathbf{x}^{(\mathbf{k+1})} = \mathbf{g}(\mathbf{x}^{(\mathbf{k})}), \text{ para } k = 0, 1, 2, \cdots$$

De forma análoga à análise da convergência deste procedimento aplicado a funções não lineares em uma variável, a convergência para a solução $\mathbf{x}^*$ é assegurada se para alguma constante $0 \leq \rho \leq 1$ ocorrer: $\left|\left|\mathbf{g}(\mathbf{x}^{(\mathbf{k})}) - \mathbf{g}(\mathbf{x}^*)\right|\right| \leq \rho \left|\left|\mathbf{x}^{(\mathbf{k})} - \mathbf{x}^*\right|\right|$, isto é, se $\mathbf{g}(\mathbf{x})$ for um mapeamento contrativo. A convergência estará garantida se $||\mathbf{g}'(\mathbf{x}^*)|| < 1$. Observe que os métodos iterativos para sistemas lineares (Seção 5.6) são casos particulares deste método, em que $\mathbf{M} = \mathbf{g}'(\mathbf{x}^*)$.

5.7.2 Método de Newton-Raphson

Como já apresentado na Seção 5.1, sistemas algébricos não lineares são representados genericamente na forma: $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, sendo $\mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ e $\mathbf{x} \in \mathbb{R}^n$.

A generalização do método de Newton-Raphson é fundamentada na *linearização* de cada uma das funções do sistema em torno do valor de seu argumento na iteração k, $\mathbf{x}^{(\mathbf{k})}$, assim:

$$f_{i,linear}(\mathbf{x}) = f_i(\mathbf{x}^{(\mathbf{k})}) + \sum_{j=1}^{n} \left.\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right|_{\mathbf{x}^{(\mathbf{k})}} (x_j - x_j^{(k)}) \text{ para } i = 1, 2, \cdots, n.$$

O valor de $\mathbf{x}^{(\mathbf{k+1})}$ é calculado de modo a anular $f_{i,linear}(\mathbf{x})$ para $i = 1, 2, \cdots, n$, assim:

$$f_i(\mathbf{x}^{(\mathbf{k})}) + \sum_{j=1}^{n} \left.\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right|_{\mathbf{x}^{(\mathbf{k})}} (x_j^{(k+1)} - x_j^{(k)}) = 0 \text{ para } i = 1, 2, \cdots, n.$$

Definindo-se $\mathbf{J}_{i,j} = \dfrac{\partial f_i(\mathbf{x})}{\partial x_j}$ como sendo o componente (i, j) da **matriz Jacobiana** de $\mathbf{f}(\mathbf{x})$ e $\mathbf{h}^{(\mathbf{k})} = \mathbf{x}^{(\mathbf{k+1})} - \mathbf{x}^{(\mathbf{k})}$ resultando no sistema algébrico linear:

$$\mathbf{f}(\mathbf{x}^{(\mathbf{k})}) + \mathbf{J}^{(\mathbf{k})}\,\mathbf{h}^{(\mathbf{k})} = \mathbf{0} \text{ ou } \mathbf{J}^{(\mathbf{k})}\,\mathbf{h}^{(\mathbf{k})} = -\mathbf{f}(\mathbf{x}^{(\mathbf{k})})$$

Dando origem ao procedimento iterativo: $\begin{cases} \mathbf{J}^{(\mathbf{k})}\,\mathbf{h}^{(\mathbf{k})} = -\mathbf{f}(\mathbf{x}^{(\mathbf{k})}) \\ \mathbf{x}^{(\mathbf{k+1})} = \mathbf{x}^{(\mathbf{k})} + \mathbf{h}^{(\mathbf{k})} \end{cases}$ para $k = 0, 1, \cdots$.

Atualmente, há inúmeros *softwares* disponíveis para obtenção analítica da matriz Jacobiana, tanto por diferenciação simbólica (MAPLE™, MATHCAD™, MATHEMATICA™, etc.) quanto por diferenciação automática que aplica a regra da cadeia (ADIFOR™, ADOL-C™, etc.).

Uma forma de reduzir o custo computacional do cálculo da matriz jacobiana é manter seu valor constante por um certo número de iterações, segundo o procedimento:

$$\mathbf{x}^{(\mathbf{k+1})} = \mathbf{x}^{(\mathbf{k})} - \alpha\,[\mathbf{J}^{(\mathbf{m})}]^{-1}\mathbf{f}(\mathbf{x}^{(\mathbf{k})}) \text{ para } k = 0, 1, \cdots$$

para $m \leq k$ e $\alpha > 0$ é um escalar que deve ser escolhido em cada iteração para minimizar o valor de $(\mathbf{f}^{\mathbf{T}}\,\mathbf{f})$. Após um certo número de iterações, quando se verifica que não há *melhora* no valor de $||\mathbf{f}||_2$, o valor de $\mathbf{J}$ deve ser atualizado. O parâmetro α, quando dentro do intervalo $(0,\,1]$, também é aplicado como uma sub-relaxação do método de Newton-Raphson, como artifício para evitar falhas de convergência.

A matiz Jacobiana pode também ser aproximada por diferenças finitas segundo um dos procedimentos: $\begin{cases} J_{i,j}(\mathbf{x^{(k)}}) \approx \dfrac{f_i(\mathbf{x^{(k)}} + \delta_{\mathbf{j}}\mathbf{I^{\langle j\rangle}}) - f_i(\mathbf{x^{(k)}})}{\delta_j} \\ J_{i,j}(\mathbf{x^{(k)}}) \approx \dfrac{f_i(\mathbf{x^{(k)}} + \delta_{\mathbf{j}}\mathbf{I^{\langle j\rangle}}) - f_i(\mathbf{x^{(k)}} - \delta_{\mathbf{j}}\mathbf{I^{\langle j\rangle}})}{2\delta_j} \end{cases}$.

Sendo $\mathbf{I^{\langle j\rangle}}$ a coluna j da matriz identidade e $\delta_j = \varepsilon_j^{abs}$ o erro absoluto para a variável x_j ou $\delta_j = \sqrt{\varepsilon}\max\{\,|x_j^{(k)}|, \varepsilon_j^{abs}, 100\sqrt{\varepsilon}\}$, sendo ε a precisão da máquina.

5.7.3 Método de Broyden

Este método busca aproximar, em cada iteração, a matriz Jacobiana, sem o cômputo das derivadas parciais das funções. O método procura estender o método da secante de funções não lineares de uma variável a funções vetoriais e foi desenvolvido originalmente por Broyden (1965)[10].

O procedimento pode ser sintetizado pelas equações abaixo:

$$\text{Início:} \begin{cases} \mathbf{x^{(0)}} \text{ (valor arbitrário)} \\ \mathbf{H_0} = \mathbf{J_0^{-1}} \\ \mathbf{x^{(1)}} = \mathbf{x^{(0)}} - \mathbf{H_0}\,\mathbf{f}(\mathbf{x^{(0)}}) \\ \delta\mathbf{x_0} = \mathbf{x^{(1)}} - \mathbf{x^{(0)}} \\ \delta\mathbf{f_0} = \mathbf{f}(\mathbf{x^{(1)}}) - \mathbf{f}(\mathbf{x^{(0)}}) \end{cases}$$

$$\text{Iteração } k = 0,\,1,\,2,\,\cdots \begin{cases} \mathbf{H_{k+1}} = \mathbf{H_k} + \dfrac{[\delta\mathbf{x_k} - \mathbf{H_k}\;\delta\mathbf{f_k}]\;(\delta\mathbf{x_k})^{\mathbf{T}}\,\mathbf{H_k}}{(\delta\mathbf{x_k})^{\mathbf{T}}\,\mathbf{H_k}\;\delta\mathbf{f_k}} \\ \mathbf{x^{(k+2)}} = \mathbf{x^{(k+1)}} - \mathbf{H_{k+1}}\,\mathbf{f}(\mathbf{x^{(k+1)}}) \\ \delta\mathbf{x_{k+1}} = \mathbf{x^{(k+2)}} - \mathbf{x^{(k+1)}} \\ \delta\mathbf{f_{k+1}} = \mathbf{f}(\mathbf{x^{(k+2)}}) - \mathbf{f}(\mathbf{x^{(k+1)}}). \end{cases}$$

A grande vantagem deste procedimento é a aproximação $\mathbf{H_k} \approx \mathbf{J^{-1}}(\mathbf{x^{(k)}})$ tornando desnecessário o cômputo das derivadas parciais de $\mathbf{f}(\mathbf{x})$ e a inversão da matriz Jacobiana, aproximando satisfatoriamente o método de Newton-Raphson.

5.7.4 Métodos de Minimização

Estes métodos consistem simplesmente na transformação do problema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, em um problema de minimização da função escalar $g(\mathbf{x}) = \sum_{i=1}^{n} f_i^2(\mathbf{x}) = \mathbf{f}(\mathbf{x})^{\mathbf{T}}\,\mathbf{f}(\mathbf{x})$ em que $g(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$. Um dos procedimentos mais empregados na minimização de funções escalares de variáveis vetoriais é o método do gradiente, conhecido como o *método da descida mais íngreme (steepest descent)*. Tal método pode ser expresso pelo procedimento recursivo:

$\mathbf{x^{(k+1)}} = \mathbf{x^{(k)}} - \alpha\nabla g(\mathbf{x^{(k)}})$ em que $\nabla g(\mathbf{x})|_j = 2\sum_{i=1}^{n} \dfrac{\partial f_i(\mathbf{x})}{\partial x_j} f_i(\mathbf{x})$, permitindo identificar:

$\nabla g(\mathbf{x}) = 2\,\mathbf{J}(\mathbf{x})^{\mathbf{T}}\,\mathbf{f}(\mathbf{x})$.

O escalar α é escolhido de modo a minimizar a função: $h(\alpha) = g\left[\mathbf{x^{(k)}} - \alpha\nabla g(\mathbf{x^{(k)}})\right]$. Uma forma simples de selecionar o valor de α é através da interpolação quadrática:

[10] Charles George Broyden (1933 –2011).

$$h(\alpha) \approx p_2(\alpha) = 2(\alpha - 1/2)(\alpha - 1)h(0) + 4\alpha(1-\alpha)h(1/2) + 2\alpha(\alpha - 1/2)h(1),$$

o valor de α que anula $p_2'(\alpha)$ é $\alpha^* = \dfrac{3h(0) - 4h(1/2) + h(1)}{4[h(0) - 2h(1/2) + h(1)]}$.

Frequentemente o método da descida mais íngreme apresenta convergência lenta e outros métodos de minimização podem ser utilizados no lugar, sendo alguns desses apresentados no Capítulo 8.

5.7.5 Homotopia e Método da Continuação

O método da continuação está baseado na variação contínua de um parâmetro na função, de modo a proporcionar condições de convergência mais fortes para os métodos de busca de raízes de funções, especialmente o método de Newton-Raphson. Quando este parâmetro representa uma combinação entre a função $\mathbf{f}(\mathbf{x})$ e outra função conhecida e de fácil solução tem-se o método da *continuação homotópica*. O tipo mais comum de homotopia é a função convexa:

$$\mathbf{h}(\mathbf{x};t) = p\,\mathbf{g}(\mathbf{x}) + (1-p)\,\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

em que $\mathbf{g}(\mathbf{x})$ é uma função de $\mathbf{x}$ de solução conhecida e o parâmetro $p \in [0,1]$.

Verificando-se que: $\mathbf{h}(\mathbf{x};1) = \mathbf{g}(\mathbf{x}) = \mathbf{0}$ e $\mathbf{h}(\mathbf{x};0) = \mathbf{f}(\mathbf{x}) = \mathbf{0}$ portanto, promovendo uma variação contínua ao parâmetro p de 1 a 0, parte-se de um ponto em que a solução é conhecida $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ em direção a uma solução de $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Pode-se afirmar que as soluções de $\mathbf{h}(\mathbf{x};p) = \mathbf{0}$ são funções do parâmetro p, representadas por $\mathbf{x}^*(p)$, sendo $\mathbf{x}^*(1)$ a solução de $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ e $\mathbf{x}^*(0)$ a solução de $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. O cálculo das raízes de $\mathbf{h}(\mathbf{x};p) = \mathbf{0}$, para cada valor fixo de p, é geralmente realizado pela aplicação do método de Newton-Raphson.

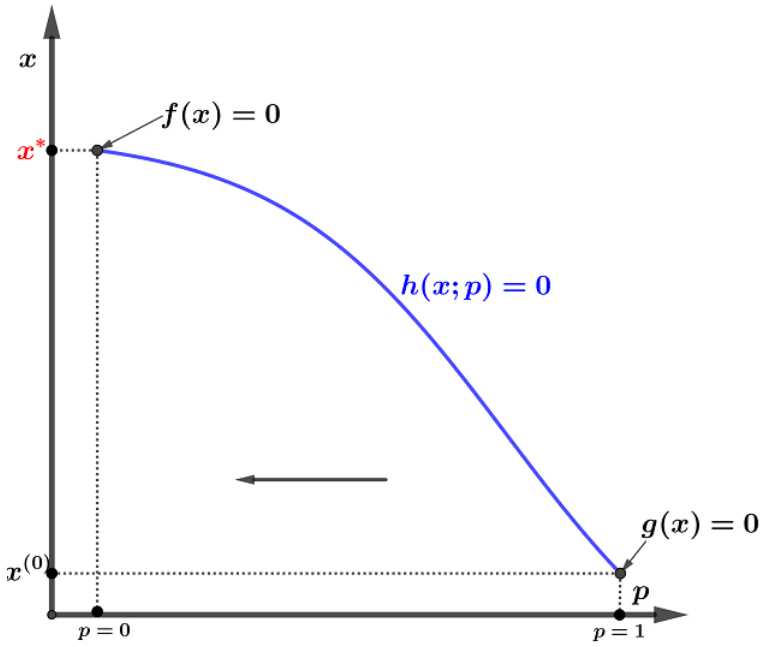


Figura 5.5: Representação simplificada do método da continuação.

As duas escolhas mais usuais da função $\mathbf{g}(\mathbf{x})$ são:

(1) **Homotopia Afim**
$\mathbf{g}(\mathbf{x}) = \mathbf{J}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)})$ que representa uma linearização de $\mathbf{f}(\mathbf{x})$ em torno de $\mathbf{x}^{(0)}$.

(2) **Homotopia de Newton**
$\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^{(0)}) \Rightarrow \mathbf{h}(\mathbf{x};\mathbf{p}) = \mathbf{f}(\mathbf{x}) - p\,\mathbf{f}(\mathbf{x}^{(0)})$.

Uma propriedade interessante do método da homotopia é que se $\mathbf{f}(\mathbf{x})$ possuir mais de uma raiz e for permitido ao parâmetro p variar sem restrição de intervalo, então é possível encontrar todas as raízes de $\mathbf{f}(\mathbf{x})$, seguindo o caminho de variação de p. Para isto, é importante que o método realize procedimentos de reparametrização quando algum componente de $\dfrac{d\mathbf{x}}{dp}$ tender a infinito, conforme descrito a seguir.

De um modo mais geral, o método da continuação consiste na solução de sistemas não lineares do tipo: $\mathbf{F}[\mathbf{x}(s),\mathbf{p}(s)] = 0$ com $\mathbf{x} \in \Re^n$ e $\mathbf{p} \in \Re^r$, em que $\mathbf{t}$ é um vetor de parâmetros e s é uma parametrização conveniente, como, por exemplo, o comprimento do arco do caminho percorrido. Considerando, por simplicidade, o caso em que $p \in \Re$, tem-se:

$$\mathbf{F}[\mathbf{x}(s), p(s)] = \mathbf{0} \Rightarrow \mathbf{J}\,\dot{\mathbf{x}}(s) + \frac{\partial \mathbf{F}}{\partial p}\dot{p}(s) = \mathbf{0},$$

sendo $\dot{\mathbf{x}}(s) = \dfrac{d\mathbf{x}}{ds}$, $\dot{p}(s) = \dfrac{dp}{ds}$ e $\mathbf{J}_{(i,j)} = \dfrac{\partial F_i}{\partial x_j}$. Caracterizando o vetor direção $\begin{pmatrix}\dot{\mathbf{x}}(s)\\ \dot{p}(s)\end{pmatrix}$ como pertencente ao espaço nulo da *Derivada de Fréchet*[11]: $D\mathbf{F} = \begin{pmatrix}\mathbf{J} & \dfrac{\partial \mathbf{F}}{\partial p}\end{pmatrix}$.

O vetor direção pode ser escalonado através de alguma normalização $N(\mathbf{x}, p, s) = 0$, o que resulta em: $\mathbf{w}^T\,\dot{\mathbf{x}}(s) + \dfrac{\partial N}{\partial p}\dot{p}(s) + \dfrac{\partial N}{\partial s} = 0$ em que $w_i = \dfrac{\partial N}{\partial x_i}$, ou seja: $\mathbf{w}^T\,\dot{\mathbf{x}}(s) + \dfrac{\partial N}{\partial p}\dot{p}(s) = -\dfrac{\partial N}{\partial s}$.

Dando origem ao sistema: $\begin{pmatrix}\mathbf{J} & \partial\mathbf{F}/\partial p\\ \mathbf{w}^T & \partial N/\partial p\end{pmatrix}\begin{pmatrix}\dot{\mathbf{x}}(s)\\ \dot{p}(s)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ -\partial N/\partial s\end{pmatrix}$.

Como exemplos de normas empregadas no escalonamento tem-se:

(a) Norma do Comprimento do Arco.

$$N_1(\mathbf{x}, t, s) = \int_{s_0}^{s} \left(||\dot{\mathbf{x}}(\xi)||^2 + \dot{p}^2(\xi)\right) d\xi - (s - s_0)$$

$w_i = \dfrac{\partial N_1}{\partial x_i} = \dot{x}_i(s)$, $\dfrac{\partial N_1}{\partial p} = \dot{p}(s)$ e $\dfrac{\partial N_1}{\partial s} = -1$ (pela aplicação do teorema da função implícita),

$$\begin{pmatrix}\mathbf{J} & \partial\mathbf{F}/\partial p\\ \dot{\mathbf{x}}^T(s) & \dot{p}(s)\end{pmatrix}\begin{pmatrix}\dot{\mathbf{x}}(s)\\ \dot{p}(s)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix}.$$

(b) Norma do Pseudo-Comprimento do Arco.

$$N_2(\mathbf{x}, t, s) = (\mathbf{x} - \mathbf{x}_0)^T\dot{\mathbf{x}}_0(s) + (p - p_0)\dot{p}_0(s) - (s - s_0)$$

$$\frac{\partial N_2}{\partial \mathbf{x}} = \dot{\mathbf{x}}_0(s),\ \frac{\partial N_2}{\partial p} = \dot{p}_0(s) \text{ e } \frac{\partial N_2}{\partial s} = -1$$

$$\begin{pmatrix}\mathbf{J} & \partial\mathbf{F}/\partial p\\ \dot{\mathbf{x}}_0^T(s) & \dot{p}_0(s)\end{pmatrix}\begin{pmatrix}\dot{\mathbf{x}}(s)\\ \dot{p}(s)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix}.$$

Esta norma é de mais simples implementação computacional.

(c) Norma de Parametrização Local (ou Interna).

$$N_3(\mathbf{x}, p, s) = \mathbf{e}_k^T\begin{pmatrix}\mathbf{x} - \mathbf{x}_0\\ p - p_0\end{pmatrix} - (s - s_0)$$

Sendo $\mathbf{e}_k$ o k-ésimo vetor unitário de dimensão $(n+1)$, resultando em:

[11] René Maurice Fréchet (1878-1973).

$\left(\mathbf{J} \quad \dfrac{\partial \mathbf{F}}{\partial p}\right)\dot{\boldsymbol{x}}(s) = \mathbf{0}$ e $\dot{x}_k = 1$, considerando $\boldsymbol{x}$ o vetor $\mathbf{x}$ estendido, isto é: $\boldsymbol{x} = \begin{pmatrix}\mathbf{x}\\ p\end{pmatrix}$.

Uma maneira de escolher o índice k é pelo critério de maior derivada direcional: k é o índice de x_i que apresenta:

$$k = \arg\max_i \left(|\dot{x}_1|, \quad |\dot{x}_2|, \quad \cdots, \quad |\dot{x}_i|, \quad \cdots, \quad |\dot{x}_n|, \quad |\dot{x}_{n+1}|\right).$$

Existem vários algoritmos para realizar continuação em caminhos regulares, alguns outros que conseguem determinar pontos limites e poucos que localizam pontos de bifurcação e continuam sobre as ramificações.

Um ponto $\begin{pmatrix}\mathbf{x}_0\\ p_0\end{pmatrix}$ é caracterizado como:

- *regular* se $\mathbf{J}(\mathbf{x}_0, p_0)$ for não singular;
- *ponto limite* se $\mathbf{J}(\mathbf{x}_0, p_0)$ for singular e $D\mathbf{F}$ tiver $posto = n$;
- *bifurcação* se $\mathbf{J}(\mathbf{x}_0, p_0)$ for singular e $D\mathbf{F}$ tiver $posto < n$.

No caso de um ponto limite, a matriz $\mathbf{J}$ pode ser tornada não singular por um procedimento de reparametrização, escolhendo, por exemplo, como novo parâmetro a variável com maior derivada direcional, possibilitando a continuação a partir deste ponto.
Métodos tipo predição-correção na continuação paramétrica são métodos que seguem os seguintes passos principais:

(1) **Obtenção de uma solução inicia**l: $(\mathbf{x_0}, p_0)$;

(2) **Predição**. Solução do sistema: $\begin{pmatrix}\mathbf{J} & \partial\mathbf{F}/\partial p\\ \mathbf{w}^T & \partial N/\partial p\end{pmatrix}\begin{pmatrix}\dot{\mathbf{x}}(s)\\ \dot{p}(s)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ -\partial N/\partial s\end{pmatrix}$.
Realizando reparametrização quando necessário e determinando o tamanho do passo $\Delta s = s - s_0$, na direção $(\mathbf{x}, p)$, para então predizer (extrapolar) $\mathbf{x}(\mathbf{s}) = \mathbf{x_0} + \dot{\mathbf{x}}(s_0)\Delta s$ e $p(s) = p_0 + \dot{p}(s_0)\Delta s$ (**método de Euler**);

(3) **Correção**. Resolução do sistema não linear $\mathbf{F}[\mathbf{x}(s), p(s)] = \mathbf{0}$ com algumas iterações do método de Newton-Raphson.

■ **Exemplo 5.5** Coluna de Destilação de uma Mistura Binária. Para comparar os diferentes métodos de resolução de sistemas algébricos não lineares considera-se a coluna de destilação de 3 pratos operada de forma contínua na destilação de uma mistura binária, apresentada na Introdução do presente Capítulo e representada no diagrama da Figura 5.1.

Considerando os seguintes dados numéricos $F = 100\,kmol/h, z = 1/2, D = 80\,kmol/h, R = 5$ e $\alpha = 4$ deseja-se calcular as vazões molares B, L e V e as composições internas da coluna.

Com os valores dos dados numéricos, chega-se ao seguinte sistema algébrico não linear:

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \left(\dfrac{24}{3x_0+1} + \dfrac{1}{4}\right)x_0 - \dfrac{25}{4}x_1 \\ -\dfrac{24x_0}{3x_0+1} + \left(\dfrac{24}{3x_1+1} + \dfrac{25}{4}\right)x_1 - \dfrac{25}{4}x_2 \\ -\dfrac{24x_1}{3x_1+1} + \left(\dfrac{24}{3x_2+1} + \dfrac{25}{4}\right)x_2 - 5x_3 \\ -\dfrac{24x_2}{3x_2+1} + \left(\dfrac{24}{3x_3+1} + 5\right)x_2 - 5x_4 \\ \dfrac{4x_3}{3x_3+1} - x_4 \end{pmatrix} - \frac{5}{8}\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}$$

A matriz Jacobiana deste sistema é:

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} \frac{24}{(3x_0+1)^2}+\frac{1}{4} & -\frac{25}{4} & 0 & 0 & 0 \\ -\frac{24}{(3x_0+1)^2} & \frac{24}{(3x_1+1)^2}+\frac{25}{4} & -\frac{25}{4} & 0 & 0 \\ 0 & -\frac{24}{(3x_1+1)^2} & \frac{24}{(3x_2+1)^2}+\frac{25}{4} & -5 & 0 \\ 0 & 0 & -\frac{24}{(3x_2+1)^2} & \frac{24}{(3x_3+1)^2}+5 & -5 \\ 0 & 0 & 0 & \frac{4}{(3x_3+1)^2} & -1 \end{pmatrix}.$$

A natureza tri-diagonal da matriz Jacobiana facilitará sua inversão e o sistema linear em cada iteração do método de Newton-Raphson pode ser resolvido pelo método de Thomas, assim:

$$\mathbf{J}(\mathbf{x^{(k)}})\,[\mathbf{x^{(k+1)}}-\mathbf{x^{(k)}}] = -\mathbf{f}(\mathbf{x^{(k)}}) \text{ para } k = 0, 1, 2, \cdots \text{ com } \mathbf{x^{(0)}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Este procedimento converge em 6 iterações para: $\mathbf{x}^* = \begin{pmatrix} 0,0088733 \\ 0,0335448 \\ 0,1173893 \\ 0,2921588 \\ 0,6227817 \end{pmatrix}$.

O mesmo problema é resolvido considerando um *congelamento* da relação de equilíbrio entre as fases resultando no método de substituições sucessivas:

$$\begin{pmatrix} \frac{24}{3x_0^{(k)}+1}+\frac{1}{4} & -\frac{25}{4} & 0 & 0 & 0 \\ -\frac{24}{3x_0^{(k)}+1} & \frac{24}{3x_1^{(k)}+1}+\frac{25}{4} & -\frac{25}{4} & 0 & 0 \\ 0 & -\frac{24}{3x_1^{(k)}+1} & \frac{24}{3x_2^{(k)}+1}+\frac{25}{4} & -5 & 0 \\ 0 & 0 & -\frac{24}{3x_2^{(k)}+1} & \frac{24}{3x_3^{(k)}+1}+5 & -5 \\ 0 & 0 & 0 & \frac{4}{3x_3^{(k)}+1} & -1 \end{pmatrix} \begin{pmatrix} x_0^{(k+1)} \\ x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \\ x_4^{(k+1)} \end{pmatrix} = \frac{5}{8} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Utiliza-se novamente a natureza tri-diagonal da matriz do sistema e o procedimento converge para a mesma solução, porém demandando 38 iterações para convergir.

O problema é novamente resolvido pelo método de Broyden, obtendo-se o mesmo resultado em 14 iterações, ressaltando-se que neste método é desnecessário o cálculo da matriz Jacobiana em cada iteração bem como a resolução do sistema linear correspondente.

Finalmente o problema é resolvido pelo método da continuação homotópica adotando-se a homotopia afim, resultando na função: $\mathbf{h}(\mathbf{x};p) = p\,\mathbf{J}(\mathbf{x^{(0)}})\,(\mathbf{x}-\mathbf{x^{(0)}})+(1-p)\,\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Diferenciando esta função em relação ao parâmetro p resulta no sistema de equações diferenciais ordinárias:

$$\frac{d\mathbf{x}(p)}{dp} = \left[p\,\mathbf{J}(\mathbf{x^{(0)}})+(1-p)\,\mathbf{J}(\mathbf{x})\right]^{-1}\,\left[\mathbf{f}(\mathbf{x})-\mathbf{J}(\mathbf{x^{(0)}})\,(\mathbf{x}-\mathbf{x^{(0)}})\right].$$

A integração é feita de $p = 1$ a $p = 0$ com a condição inicial: $\mathbf{x}(\mathbf{1}) = \mathbf{x}^{(\mathbf{0})}$. Tal sistema é integrado pelo método de Runge-Kutta de quarta ordem com passo fixo, resultando no final ($p = 0$) nos mesmos resultados anteriores, a evolução do procedimento é apresentada na Figura 5.6.

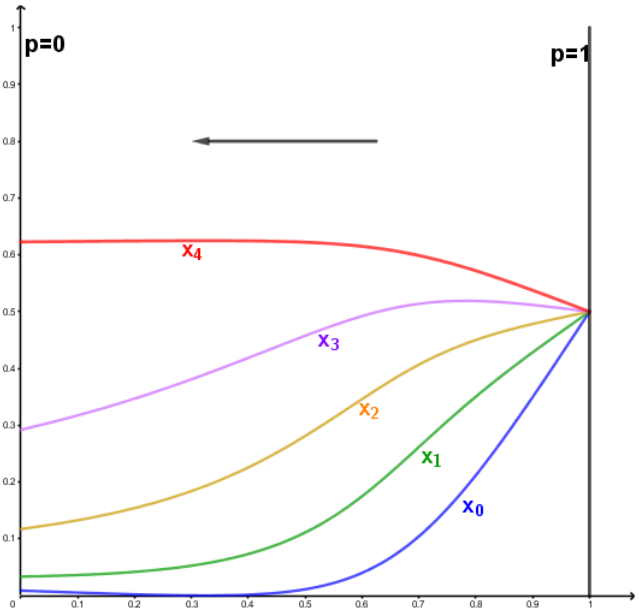


Figura 5.6: Evolução do caminho homotópico.

Conforme apresentado na Seção 5.1, um novo problema se configura com a especificação de F, z, x_D, e x_B Considerando os valores numéricos: $F = 100\,kmol/h$, $z = 1/2$, $x_D = 0,8$, e $x_B = 0,02$, tem-se $x_4 = x_D = \frac{4}{5}$, $x_0 = x_B = \frac{1}{50}$ e as vazões B e D podem ser calculadas pelos balanços globais:

$$\begin{cases} B + D = 100 \\ B/50 + 4D/5 = 50 \end{cases} \Rightarrow B = \frac{500}{13},\ D = \frac{800}{13} \text{ e } \frac{F}{D} = \frac{13}{8},$$ e as vazões molares L e V podem ser expressas pela razão de refluxo R: $L = \frac{800}{13}R$ e $V = \frac{800}{13}(1+R)$. A composição do vapor no refervedor é calculada por: $y_0 = \frac{4x_0}{3x_0+1} = \frac{4}{53}$ e $y_0 - x_0 = \frac{147}{2650}$ e pelo balanço no condensador tem-se: $y_3 = x_4 = \frac{4}{5} \Rightarrow x_3 = \frac{y_3}{4-3y_3} = \frac{1}{2}$.

Da equação de balanço do refervedor obtém-se: $x_1 = \frac{1}{50} + \frac{147}{2650}\frac{(1+R)}{R+13/8} = x_1(R)$ e da equação de balanço do prato 3: $y_2 = \frac{4}{5} - \frac{3}{10}\frac{R}{1+R} = y_2(R)$. A substituição destas duas expressões na equação de balanço molar do prato 1 dá origem a:

$$\Phi(R) = (1+R)\left(\frac{4x_1(R)}{3x_1(R)+1} - \frac{4}{53}\right) + (R+13/8)\left(x_1(R) - \frac{y_2(R)}{4-3y_2(R)}\right) = 0.$$

Esta última equação foi resolvida pelos procedimentos usuais de resolução de equações algébricas não lineares em uma variável resultando em $R = 7,9581$, determinando-se a seguir: $L = 489,7321\,kmol/h$, $V = 551,2706\,kmol/h$ e

i	0	1	2	3	4
x_i	0,0200	0,0719	0,2223	0,5000	0,8000
y_i	0,0755	0,2364	0,5335	0,8000	

■

■ **Exemplo 5.6** Reator do Gás de Síntese, Exemplo 5.5 de Carnahan, Luther e Wilkes (1969). As principais reações que ocorrem na produção de gás de síntese através da oxidação parcial do metano com oxigênio são:

$$\begin{cases} CH_4 + 1/2O_2 \leftrightarrows CO + 2H_2 \\ CH_4 + H_2O \leftrightarrows CO + 3H_2 \\ H_2 + CO_2 \leftrightarrows CO + H_2O \end{cases}$$

Deseja-se calcular a relação entre as vazões molares de oxigênio e metano na alimentação de um reator de gás de síntese operando adiabaticamente, de modo que a temperatura de equilíbrio no interior do reator seja igual a $2200°F$. A pressão de operação do reator é igual a $20atm$ e a temperatura de entrada dos reagentes é igual a $1000°F$.

Considerando o comportamento da mistura reacional como ideal as seguintes relações de equilíbrio prevalecem:

$$\begin{cases} \text{Reação 1: } K_1 = \dfrac{p_{CO}\, p_{H_2}^2}{p_{CH_4}\, p_{O_2}^{1/2}} = 1,3 \times 10^{11} \\ \text{Reação 2: } K_2 = \dfrac{p_{CO}\, p_{H_2}^3}{p_{CH_4}\, p_{H_2O}} = 1,7837 \times 10^5 \\ \text{Reação 3 : } K_3 = \dfrac{p_{CO}\, p_{H_2O}}{p_{CO_2}\, p_{H_2}} = 2,6058 \end{cases}$$

Em que $p_{CO}, p_{CO_2}, p_{H_2O}, p_{H_2}, p_{CH_4}$ e p_{O_2} são as pressões parciais, respectivamente, de $CO, CO_2, H_2O, H_2, CH_4$ e O_2.

As entalpias dos componentes presentes no processo a 1000 e $2200°F$ são tabeladas a seguir:

Componente	$H(1000°F)$ $(BTU/lbmol)$	$H(2200°F)$ $(BTU/lbmol)$
CO	-38528	-28837
CO_2	-154958	-139009
H_2O	-90546	-78213
H_2	10100	18927
CH_4	-13492	8427
O_2	10690	20831

Uma quarta reação também ocorre a altas temperaturas:

Reação 4: $C_{solido} + CO_2 \leftrightarrows 2CO$ com $K_4 = \dfrac{p_{CO}^2}{a_C\, p_{CO_2}} \equiv 1329,5$ sendo a_C a atividade do carbono no estado sólido (seu valor pode ser considerado como unitário).

Considerando de início a inexistência da reação 4 e as seguintes variáveis:

x_1 : número de mols de CO no equilíbrio/mol de CH_4 na alimentação;
x_2: número de mols de CO_2 no equilíbrio/mol de CH_4 na alimentação;
x_3 : número de mols de H_2O no equilíbrio/mol de CH_4 na alimentação;
x_4 : número de mols de H_2 no equilíbrio /mol de CH_4 na alimentação;
x_5: número de mols de CH_4 no equilíbrio /mol de CH_4 na alimentação;
x_6: número de mols de O_2 na alimentação /mol de CH_4 na alimentação;
x_7: número total de mols dos produtos /mol de CH_4 na alimentação.

Devido ao valor elevado da constante de equilíbrio da primeira reação, pode-se considerar como se todo o oxigênio alimentado ao sistema fosse consumido, isto é: $p_{O_2} \cong 0$. Com esta consideração os balanços de massa de cada elemento químico presente e o balanço de energia conduzem ao seguinte sistema de equações:

$$\begin{cases} \text{Balanço de oxigênio: } 2x_6 = x_1 + 2x_2 + x_3 \\ \text{Balanço de hidrogênio: } 4 = 2x_3 + 2x_4 + 4x_5 \\ \text{Balanço de carbono: } 1 = x_1 + x_2 + x_5 \\ \text{Balanço global do produto: } x_7 = x_1 + x_2 + x_3 + x_4 + x_5 \\ \text{Equilíbrio da reação 2: } P_{total}^2 x_1 x_4^3 = 1,7837 \times 10^5 x_3 x_5 x_7^2 \\ \text{Equilíbrio da reação 3: } x_1 x_3 = 2,6058 x_2 x_4 \\ \text{Balanço de energia: } -28837x_1 - 139009x_2 - 78213x_3 + 18927x_4 + 8427x_5 = -13492 + 10690x_6 \end{cases}$$

Resolver este sistema de equações algébricas. Após resolver o sistema calcule; $\bar{K} = \dfrac{p_{CO}^2}{a_C\, p_{CO_2}} \cong \dfrac{P_{total}\, x_1^2}{x_2\, x_7}$ se $\bar{K} > K_4 = 1329,5$ há a possibilidade de carvão sólido no interior do reator, caso contrário tal não ocorre. Verificar qual das possibilidades prevalece.

Refazer o problema considerando a possibilidade da reação 1 não ser completa, o que implica na existência de oxigênio residual no produto. Introduzindo o parâmetro $\beta = \dfrac{P_{total}^2}{1,7837 \times 10^5} = 2,2425 \times 10^{-3}$ e reescalando o balanço de energia dividindo ambos os membros por 139009, chega-se ao seguinte sistema algébrico não linear:

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \dfrac{x_1}{2} + x_2 + \dfrac{x_3}{2} - x_6 \\ \dfrac{x_3}{2} + \dfrac{x_4}{2} + x_5 - 1 \\ x_1 + x_2 + x_5 - 1 \\ -0,2074x_1 - x_2 - 0,5626x_3 + 0,1362x_4 + 0,0606x_5 - 0,0769x_6 + 0.0971 \\ x_1 x_3 - 2,6058 x_2 x_4 \\ 2,2425 \times 10^{-3} x_1 x_4^3 - x_3 x_5 (x_1 + x_2 + x_3 + x_4 + x_5)^2 \end{pmatrix}$$

A matriz Jacobiana deste sistema é:

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} 1/2 & 1 & 1/2 & 0 & 0 & -1 \\ 0 & 0 & 1/2 & 1/2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ -0,2074 & -1 & -0,5626 & 0,1362 & 0,0606 & -0,0769 \\ x_3 & -2,6058x_4 & x_1 & -2,6058x_2 & 0 & 0 \\ g_1(\mathbf{x}) & g_2(\mathbf{x}) & g_3(\mathbf{x}) & g_4(\mathbf{x}) & g_5(\mathbf{x}) & 0 \end{pmatrix}$$

Em que:
$$\begin{cases} g_1(\mathbf{x}) = 2,2425 \times 10^{-3} x_4^3 - 2x_3 x_5 (x_1 + x_2 + x_3 + x_4 + x_5) \\ g_2(\mathbf{x}) = -2x_3 x_5 (x_1 + x_2 + x_3 + x_4 + x_5) \\ g_3(\mathbf{x}) = -x_5 (x_1 + x_2 + x_3 + x_4 + x_5)[(x_1 + x_2 + x_3 + x_4 + x_5) + 2x_3] \\ g_4(\mathbf{x}) = 6,7276 \times 10^{-3} x_1 x_4^2 - 2x_3 x_5 (x_1 + x_2 + x_3 + x_4 + x_5) \\ g_5(\mathbf{x}) = -x_3 (x_1 + x_2 + x_3 + x_4 + x_5)[(x_1 + x_2 + x_3 + x_4 + x_5) + 2x_5]. \end{cases}$$

Aplicando o método de Newton-Raphson adotando a condição inicial (em que se considera apenas a ocorrência completa da Reação 1): $(\mathbf{x}^{(\mathbf{0})})^T = \begin{pmatrix}1 & 0 & 0 & 2 & 0 & 1/2\end{pmatrix}$, o procedimento numérico converge após 6 iterações para:

$$(\mathbf{x}^{(6)})^T = \begin{pmatrix}0,961465 & 0,027466 & 0,137033 & 1,840831 & 0,011068 & 0,576715\end{pmatrix}.$$

Com estes valores calcula-se o número total de mols da mistura: $n_{total} = x_1 + x_2 + x_3 + x_4 + x_5 = 2,9779$ e, a seguir, as frações molares da mistura:

$X_1 = 0,3229$: fração molar do CO no equilíbrio;
$X_2 = 0,0092$: fração molar de CO_2 no equilíbrio;
$X_3 = 0,046$: fração molar de H_2O no equilíbrio;
$X_4 = 0,6182$: fração molar de H_2 no equilíbrio;
$X_5 = 0,0037$: fração molar de CH_4 no equilíbrio.

Verificação da possibilidade de existência de carvão sólido: $\bar{K} \cong \dfrac{P_{total} x_1^2}{x_2 x_7} = 226,0424 < K_4 = 1329,5$ constando-se que não há carbono sólido presente no reator!

O problema foi novamente resolvido pelo método de Broyden que converge ao mesmo resultado com 8 iterações. O método da continuação, com a homotopia afim, foi também aplicado chagando-se novamente aos mesmos valores finais.

A resolução do problema considerando a possibilidade da reação 1 não ser completa, o que implica na existência de oxigênio residual no produto, aumenta a dimensão do sistema e a inclusão de uma nova variável que é o número de mols de O_2 no equilíbrio. O problema é então novamente resolvido adotando os valores das demais variáveis iguais aos valores obtidos anteriormente e o número de mols de O_2 no equilíbrio igual a zero, os novos valores obtidos em pouco diferem dos valores anteriores e o número de mols de O_2 no equilíbrio obtido é igual a $2,2576 \times 10^{-6}$. ∎

5.8 Problemas Propostos

Problema 5.1 Refaça o Exemplo 5.5 utilizando a linearidade das quatro primeiras equações permitindo expressar as variáveis x_1, x_2, x_3 e x_4 em função de x_5 e x_6. Substituindo, a seguir, estas expressões nas duas últimas equações o sistema algébrico reduz-se a duas equações de x_5 e x_6. Resolva o novo sistema aplicando um método conveniente. Compare os resultados obtidos com os anteriores e discuta as prováveis vantagens e desvantagem do novo procedimento.

Problema 5.2 Resolva a equação de diferenças:

$$u_{i+2} + 4\,i\,u_{i+1} + u_i = i \text{ para } i = 1, 2 \cdots, n \text{ com } u_1 = 0 \text{ e } u_{n+2} = 1$$

Resolva para n = 6, 7 e 8.

Refaça o problema com as condições de contorno: $u_1 = u_{n+2}$ e $u_{n+2} = 2u_{n+1}$.

Problema 5.3 Resolva a equação de diferenças não linear:

$$u_{i+2} + 4\,i\,u_{i+1} + u_i^2 = (i-1)(10-i) \text{ para } i = 1, 2 \cdots, 10 \text{ com } u_1 = 0,5 \text{ e } u_{12} = -2$$

Problema 5.4 Cinco reatores de mistura iguais conduzem, isotermicamente e em fase líquida, uma reação irreversível de segunda ordem. Deseja-se determinar a concentração do reagente na saída do quinto reator nas condições estacionárias de operação do processo, para isto deve-se resolver o sistema de equações algébricas (que traduzem os balanços molares do reagente em cada reator escritos em forma adimensional):

$$\begin{cases} \text{Balanço molar no primeiro reator: } x_1(1+\alpha x_1) = 1; \\ \text{Balanço molar no ssegundo reator: } x_2(1+\alpha x_2) = x_1; \\ \text{Balanço molar no terceiro reator: } x_3(1+\alpha x_3) = x_2; \\ \text{Balanço molar no quarto reator: } x_4(1+\alpha x_4) = x_3; \\ \text{Balanço molar no quinto reator: } x_5(1+\alpha x_5) = x_4. \end{cases}$$

Considerando $\alpha = 1,5$ determine a concentração do reagente na saída de cada um dos reatores. Utilize em sua resolução a natureza bi-diagonal da matriz Jacobiana. Investigue a possibilidade de transformar o problema na resolução de uma equação não linear em uma variável, o que pode ser feito expressando a concentração de entrada em cada um dos reatores em função da variável de saída e buscando o valor de x_5 que *conduz* à concentração de entrada do primeiro reator ao valor unitário.

Problema 5.5 O modelo estacionário do estágio i de uma coluna de absorção de pratos é descrito pela equação de balanço de massa:

$$Lx_{i+1} + Vy_{i-1} = Lx_i + V\,y_i \text{ para } i = 1, 2, \cdots, N.$$

Sendo:
$$\begin{cases} N: \text{ número total de pratos;} \\ L: \text{ vazão molar da fase líquida;} \\ V: \text{ vazão molar da fase gasosa;} \\ x_i: \text{ fração molar do soluto na fase líquida;} \\ y_i: \text{ fração molar do soluto na fase gasosa.} \end{cases}$$

Sabendo-se que a relação de equilíbrio entre as fases é linear e dada pela expressão: $y_i = mx_i$, sugira um procedimento iterativo para resolver este sistema conhecendo-se: L, V, m, y_0 e x_{N+1}. Para ilustrar seu procedimento adote: $L = 40\,kgmol/h$, $V = 65\,kgmol/h$, $m = 0,75$, $N = 10$, $y_0 = 0,25$ e $x_1 1 = 0.$. Avalie o que ocorre com as composições de saída, x_1 e y_N, quando N tende ao infinito.

Refaça o problema com a relação de equilíbrio não linear $y_i = \dfrac{mx_i}{1+\alpha x_i}$ com , $m = 0,75$ e $\alpha = 0,05$.

Problema 5.6 O modelo estacionário do estágio i de uma coluna de absorção de pratos, na qual ocorre uma reação química irreversível na fase líquida, é descrito pela equação de balanço de massa:

$$Lx_{i+1} + V\,y_{i-1} = Lx_i + Vy_i + H\,k\,x_i^2 \text{ para } i = 1, 2, \cdots, N.$$

Sendo:
$$\begin{cases} N = 12: \text{ número total de pratos;} \\ L = 40 kgmol/h: \text{ vazão molar da fase líquida;} \\ V = 60 kgmol/h: \text{ vazão molar da fase gasosa;} \\ H = 20 kgmol: \text{ número de mols da fase líquida no prato } i; \\ k = 0,5 h^{-1}: \text{ constante de velocidade da reação;} \\ x_i: \text{ fração molar do soluto na fase líquida;} \\ y_i: \text{ fração molar do soluto na fase gasosa.} \end{cases}$$

Sabendo-se que a relação de equilíbrio entre as fases é dada pela expressão: $y_i = \dfrac{mx_i}{1+\alpha x_i}$ com $m = 0,75$ e $\alpha = 0,05$ e que $y_0 = 0,45$ e $x_{13} = 0$ (alimentação da fase líquida isenta de soluto). Determine as composições do soluto em todos os pratos em ambas as fases.

Problema 5.7 Em um conjunto de reações químicas, para determinar o número de reações independentes monta-se uma matriz composta pelos coeficientes estequiométricos das reações

considerando-os como positivo quando o componente for reagente na reação correspondente e como negativo quando o componente for produto na reação (esta matriz se chama de *matriz estequiométrica*). Assim para o conjunto de reações químicas:

$$\begin{cases} \text{Reação 1: } 4NH_3 + 5O_2 \leftrightarrows 4NO + 6H_2O \\ \text{Reação 2: } 4NH_3 + 3O_2 \leftrightarrows 2N_2 + 6H_2O \\ \text{Reação 3: } 4NH_3 + 6NO \leftrightarrows 5N_2 + 6H_2O \\ \text{Reação 4: } 2NO + O_2 \leftrightarrows 2NO_2 \\ \text{Reação 5: } 2NO \leftrightarrows N_2 + O_2 \\ \text{Reação 6: } N_2 + 2O_2 \leftrightarrows 2NO_2 \end{cases}$$

A *matriz estequiométrica* correspondente a este esquema de reações é:

$$\mathbf{A} = \begin{pmatrix} 4 & 5 & -6 & 0 & -4 & 0 \\ 4 & 3 & -6 & -2 & 0 & 0 \\ 4 & 0 & -6 & -5 & 6 & 0 \\ 0 & 1 & 0 & 0 & 2 & -2 \\ 0 & -1 & 0 & -1 & 2 & 0 \\ 0 & 2 & 0 & 1 & 0 & -2 \end{pmatrix}.$$

O número de reações independentes é igual ao posto da matriz **A**, sendo neste exemplo igual a três. Baseado nesta informação indique três reações do esquema apresentado que sejam independentes entre si, justificando sua escolha pelo cálculo do posto da matriz estequiométrica das reações escolhidas.

Problema 5.8 Conforme apresentado na Seção 5.1, as equações do sistema hidráulico da Figura 5.2 podem ser expressas em formas adimensionais por:

$$\begin{cases} f_1(x_1,x_2) = x_1 - 1 - \alpha + \beta (x_2)^{\frac{3}{2}}; \\ f_2(x_1,x_2) = x_1 - 1 - \gamma (x_2)^2. \end{cases}$$

Sendo: $x_1 = \dfrac{p_2}{p_1}$, $x_2 = \dfrac{Q}{Q_{ref}}$, $\alpha = \dfrac{a}{p_1}$, $\beta = \dfrac{b\,(Q_{ref})^{\frac{3}{2}}}{p_1}$ e $\gamma = 8\dfrac{f_M\,\rho\,L\,(Q_{ref})^2}{\pi^2 D^5 p_1}$, variáveis e parâmetros adimensionais. Sendo: $p_1 = 1\,atm = 1{,}013 \times 10^5\,Pa$.

	Dados 1	Dados 2
$Q_{ref}\left(\dfrac{m^3}{s}\right)$	$0{,}720 \times 10^{-3}$	$5{,}456 \times 10^{-3}$
α	1,136	2,62
β	0,136	1,62
γ	0,150	0,337

Adotando com valor inicial do processo iterativo, para ambos os dados, $\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, determine x_1 e x_2 para os dois conjuntos de dados.

Na realidade o fator de atrito de Moody no interior da tubulação é função do número de Reynolds, $Re = \dfrac{D\,\rho\,\bar{u}}{\mu}$ em que $\bar{u} = \dfrac{Q}{\pi(D/2)^2}$ (velocidade média no interior da tubulação), e μ (viscosidade do líquido: $\mu_1 = 1{,}005$ centipoise e $\mu_2 = 2{,}46$ centipoise) e da rugosidade interna do tubo, ε. Esta dependência é expressa por:

$$\begin{cases} f_M = \dfrac{64}{Re} \text{ se } Re \leq 2000; \\ \dfrac{1}{\sqrt{f_M}} = -2\log\left(\dfrac{\varepsilon}{3{,}7D} + \dfrac{2{,}51}{Re\sqrt{f_M}}\right) \text{ se } Re > 2000 \text{ (\textit{Equação de Colenbrook}).} \end{cases}$$

Um bom chute inicial para a resolução numérica da equação de Colebrook é a equação de Blassius expressa por $f_M = 0,316Re^{-0,25}$ que é válida para tubulações lisas ($\varepsilon = 0$) e escoamento turbulento. Refaça o problema para os dois conjuntos de dados e com estas novas considerações usando $\varepsilon = 0,0005\,ft$.

Problema 5.9 Considere o sistema hidráulico esquematizado na Figura 5.7.

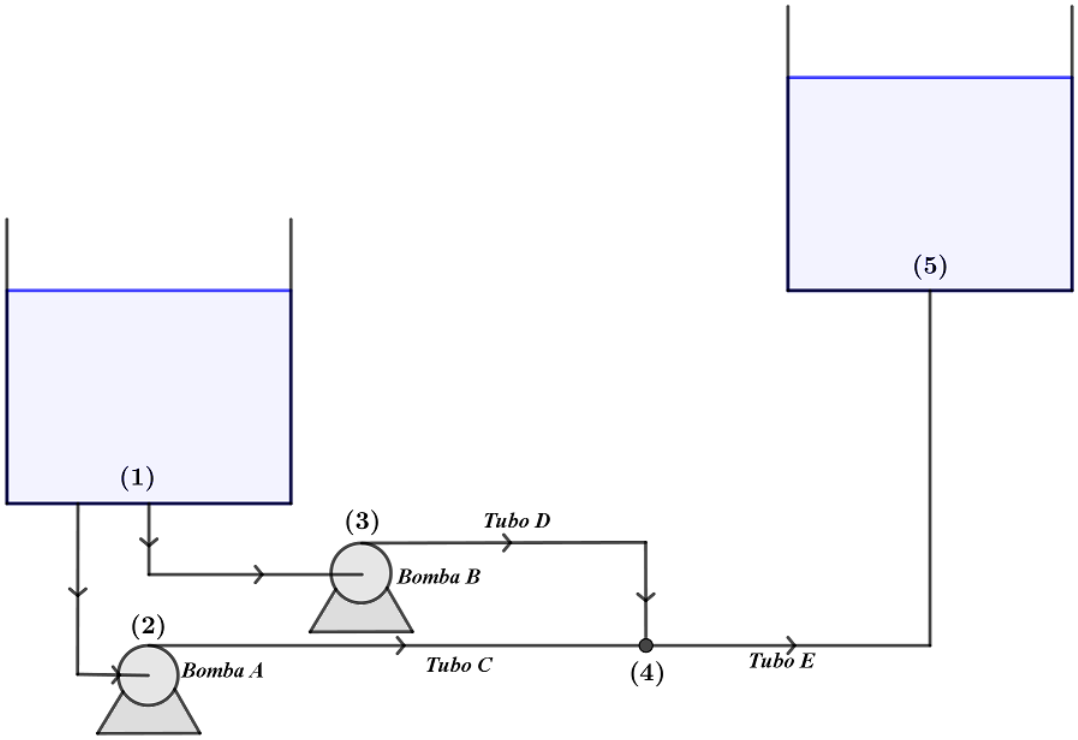


Figura 5.7: Sistema hidráulico do Problema 5.9.

As pressões p_1 e p_5 nos pontos **(1)** e **(5)** podem ser consideradas como iguais à pressão atmosférica. As equações que descrevem o escoamento em cada trecho do sistema são:

$$\begin{cases} \text{Ponto de junção (4): } Q_E = Q_D + Q_C; \\ \text{Bomba A: } p_2 - p_1 = \alpha_A - \beta_A Q_C^2; \\ \text{Bomba B: } p_3 - p_1 = \alpha_B - \beta_B Q_D^2; \\ \text{Perda de carga no tubo C: } p_2 - p_4 = 8\dfrac{f_M \rho L_C Q_C^2}{\pi^2 D_C^5}; \\ \text{Perda de carga no tubo D: } p_3 - p_4 = 8\dfrac{f_M \rho L_D Q_D^2}{\pi^2 D_D^5}; \\ \text{Perda de carga no tubo E: } p_4 - p_5 + \rho g(z_5 - z_4) = 8\dfrac{f_M \rho L_E Q_E^2}{\pi^2 D_E^5}. \end{cases}$$

Em que: $(z_5 - z_4) = 70\,ft$ (elevação da tubulação), $f_M = 0,02792$, $\rho = 62,43\,lb_m/ft^3$ (massa específica da água), $p_1 = p_5 = 1,00\,atm$ e

Bomba	α (psia)	β $\frac{psi}{(gpm)^2}$
A	156,6	0,00752
B	117,1	0,00427

Tubulação	D (polegadas)	L (pés)
C	1,278	125
D	2,067	125
E	2,469	145

Calcule p_2, p_3, p_4, Q_C, Q_D e Q_E. (sugestão: Adote o mesmo tipo de adimensionamento do Problema 5.8).

Problema 5.10 O seguinte conjunto de pontos do plano (x_i, y_i) para $i = 0, 1, \cdots, 10$, está disponível:

i	0	1	2	3	4	5	6	7	8	9	10
x_i	0	0,877	2,034	2,940	4,129	4,870	6,084	6,922	7,837	8,859	10
y_i	0,369	0,719	0,982	0,930	0,688	0,683	0,863	0,881	0,517	0,632	0,336

Desenvolva a função *spline* de terceiro grau apropriada para este conjunto, após esta determinação interpole os valores de y para os valores inteiro de x no intervalo [0,10].

6. Integração Numérica

6.1 Introdução

Vimos nos Capítulos 2 e 3 que entre os motivos para o uso de polinômios na aproximação de funções está a facilidade de cálculos de derivadas e integrais. Neste capítulo aplicaremos as aproximações polinomiais para a integração numérica de funções, ou seja:

$$I = \int_a^b f(x)dx \approx \int_a^b p_n(x)dx.$$

Esta aproximação, quando escrita na forma:

$$I = \int_a^b f(x)dx \approx \sum_{i=0}^{n} \omega_i f(x_i).$$

é chamada de **quadratura numérica**, tal nomenclatura advém do termo *quadratura do círculo* proposto pelos antigos matemáticos gregos (aproximadamente cinco séculos a.C.) que consistia em construir um quadrado com a mesma área de um dado círculo.

Adotando o polinômio interpolador de Lagrange para representar $p_n(x)$:

$$p_n(x) = \sum_{i=0}^{n} \ell_i(x) f(x_i),$$

e sabendo que o erro de truncamento da aproximação de $f(x) = p_n(x) + R_n(x)$ é dado por:

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \text{ com } \xi \in (a, b),$$

então a integral de $f(x)$ no intervalo $[a, b]$ pode ser escrita da seguinte forma:

$$I = \int_a^b f(x)dx = \int_a^b \sum_{i=0}^{n} \ell_i(x) f(x_i)dx + \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)dx.$$

e uma aproximação para o cálculo desta integral é:

$$I = \int_a^b f(x)dx \approx \sum_{i=0}^{n} \left(\int_a^b \ell_i(x)dx \right) f(x_i) = \sum_{i=0}^{n} \omega_i f(x_i).$$

em que $\omega_i = \int_a^b \ell_i(x)dx$ e o erro dessa aproximação da integral é:

$$Erro_n = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n}(x-x_i)dx \text{ com } \xi \in (a,b).$$

A determinação dos valores de ω_i, denominados *pesos da quadratura*, é bastante facilitada se o intervalo $[a,b]$ for normalizado por um dos procedimentos descritos Capítulo 3.

- Mudança de variável para $t = \dfrac{2x-(a+b)}{b-a} \Rightarrow dx = \left(\dfrac{b-a}{2}\right) dt$ resultando em:

$$I = \left(\frac{b-a}{2}\right) \int_{-1}^{+1} f[x(t)]dt \approx \left(\frac{b-a}{2}\right) \sum_{i=0}^{n} \omega_i f[x(t_i)],$$

  em que $x(t_i) = \left(\dfrac{1-t_i}{2}\right) a + \left(\dfrac{1+t_i}{2}\right) b.$

- Mudança de variável para $t = \dfrac{x-a}{b-a} \Rightarrow dx = (b-a)dt$ resultando em:

$$I = (b-a) \int_0^{+1} f[x(t)]dt \approx (b-a) \sum_{i=0}^{n} \omega_i f[x(t_i)],$$

  em que $x(t_i) = (1-t_i)\,a + t_i\,b.$

6.2 Método de Integração Numérica de Newton-Cotes

O método de integração numérica de Newton-Cotes[1] é desenvolvido a partir da aproximação do integrando por uma função polinomial de grau n obtida pela interpolação polinomial de mesmo grau, usando $(n+1)$ pontos nodais igualmente espaçados no interior do intervalo $[a,b]$, ou seja, $x_i = x_0 + ih$ para $i = 0, 1, 2, \cdots, n$. Essas fórmulas são ditas **fórmulas fechadas** quando $x_0 = a$ e $x_n = b$, com $h = (b-a)/n$, e **fórmulas abertas** quando x_0 e x_n estão dentro do intervalo $[a,b]$, com $h = (b-a)/(n+2)$.

Assim, com $x_i = x_0 + ih$ para $i = 0, 1, \cdots, n$, estes pontos nodais correspondem a $t_i = \dfrac{i}{n}$ para as fórmulas fechadas e $t_i = \dfrac{i+1}{n+2}$ para as fórmulas abertas, no reescalamento $\dfrac{x-a}{b-a}$.

As fórmulas fechadas para $n = 1$ e $n = 2$ também são conhecidas como regra ou **fórmula do trapézio** e **fórmula de Simpson**[2], respectivamente. A fórmula aberta para $n = 0$ é conhecida como **fórmula do ponto médio** (ou **fórmula do retângulo**) e para $n = 1$ também é chamada de fórmula do trapézio, mas com pontos nodais diferentes.

Tomando como exemplo a obtenção da fórmula do trapézio, com $n = 1, x_0 = a, x_1 = b$ ou $t_0 = 0, t_1 = 1$ e $h = b-a$, o polinômio interpolador $p_1(t)$ é dado por:

$$p_1(t) = \frac{t-t_1}{t_0-t_1} f(x_0) + \frac{t-t_0}{t_1-t_0} f(x_1) = (1-t)f(x_0) + t\,f(x_1),$$

[1] Roger Cotes (1682-1716).
[2] Thomas Simpson (1710-1761).

resultando para o cálculo da integral:

$$I = (b-a)\int_0^1 f[x(t)]dt \approx h\int_0^1 p_1(t)dt = h\frac{[f(x_0)+f(x_1)]}{2},$$

que é igual à área do trapézio de base h e altura média $[f(x_0)+f(x_1)]/2$, na qual também identificam-se $\omega_0 = \omega_1 = 1/2$. O erro desta aproximação é dado por:

$$Erro_1 = \frac{1}{2!}\int_a^b f''[\xi(x)](x-x_0)(x-x_1)dx = \frac{h^3}{2!}\int_0^1 f''[\xi(x_0+ht)]\,t\,(t-1)dt.$$

Como $t\,(t-1)$ não muda de sinal no intervalo $(0,\,1)$, o teorema do valor médio da integral pode ser aplicado:

$$Erro_1 = \frac{h^3}{2!}\int_0^1 f''[\xi(x_0+ht)]\,t\,(t-1)dt = \frac{h^3}{2!}f''(\xi)\int_0^1 t\,(t-1)dt = -\frac{h^3 f''(\xi)}{12} \text{ com } \xi \in (a,b).$$

Uma forma mais simples de determinar os *pesos da quadratura* é considerando as integrais: $\int_0^1 t^k dt = \frac{1}{k+1} = \sum_{i=0}^{n} \omega_i t_i^k$ para $k = 0,\,1,\,\cdots,\,n,$ resultando no sistema algébrico linear (para as fórmulas fechadas, $t_i = i/n$):

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1/n & 2/n & \cdots & 1 \\ 0 & (1/n)^2 & (2/n)^2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & (1/n)^n & (2/n)^n & \cdots & 1 \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ \vdots \\ 1/(n+1) \end{pmatrix}.$$

Para o exemplo anterior, com $n = 1$, o sistema linear fica:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix},$$

cuja solução é $\omega_0 = \omega_1 = 1/2$.

Verificando-se que a integração numérica de uma função polinomial em t de grau m é exata para $m \leq n$, expressando esse polinômio pelo polinômio interpolador da Lagrange com os $(n+1)$ pontos nodais $t_i = \dfrac{i}{n}$ para $i = 0,\,1,\,\cdots,\,n$.

$$p_m(t) = \sum_{j=0}^{n} \ell_j(t)p_m(t_j) + R_n(t),$$

sendo o erro da aproximação:

$$R_n(t) = \begin{cases} 0 \text{ se } m \leq n \\ p_{nodal}(t)q_{m-(n+1)}(t) \text{ se } m > n \end{cases},$$

em que $p_{nodal}(t) = \prod_{i=0}^{n}(t-i/n)$ e $q_{m-(n+1)}(t)$ é um polinômio em t de grau $m-(n+1)$.

O polinômio nodal é um polinômio de grau $(n+1)$ cujas raízes se distribuem simetricamente em torno de $t = 1/2$, assim reescrevendo $p_{nodal}(t)$ em termos de $z = t - 1/2$ verifica-se que $p_{nodal}(z) = \prod_{i=0}^{n}\left(z + \frac{1}{2} - \frac{i}{n}\right)$ apresenta as raízes dispostas simetricamente em torno de $z = 0$ e

$z=0$ só será raiz se n for um número par. Desse modo, se n for um número par $p_{nodal}(z)$ só contém potências ímpares de z e se n for um número ímpar $p_{nodal}(z)$ só contém potências pares de z, o que implica em: $\int_0^1 p_{nodal}(t)dt = \int_{-1/2}^{+1/2} p_{nodal}(z)dz = 0$ se n for par, garantindo que a integral $\int_0^1 p_m(t)dt = \sum_{i=0}^{n} \omega_i p_m(t_i)$ é exata para $m \leq (n+1)$ caso n for par (um número par de sub-intervalos ou um número ímpar de pontos nodais). Outra consequência da distribuição das raízes de $p_{nodal}(t)$ é a aplicação do teorema do valor médio de acordo com:

$$\begin{cases} \int_0^1 p_{nodal}(t)f(t)dt = f(\bar{t})\int_0^1 p_{nodal}(t)dt \text{ se } n \text{ for ímpar} \\ \int_0^1 (t-1/2)p_{nodal}(t)f(t)dt = f(\bar{t})\int_0^1 (t-1/2)p_{nodal}(t)dt \text{ se } n \text{ for par} \end{cases}.$$

Baseado nestas característica do $p_{nodal}(t)$ pode-se calcular o erro da integração numérica do método de Newton-Cotes segundo:

$$Erro = \begin{cases} \dfrac{(b-a)^{n+2}}{(n+1)!}\left.\dfrac{d^{n+1}f(x)}{dx^{n+1}}\right|_{x=\xi}\int_0^1 p_{nodal}(t)dt \text{ se } n \text{ for ímpar} \\ \dfrac{(b-a)^{n+3}}{(n+2)!}\left.\dfrac{d^{n+2}f(x)}{dx^{n+2}}\right|_{x=\xi}\int_0^1 (t-1/2)p_{nodal}(t)dt \text{ se } n \text{ for par} \end{cases}.$$

Após a determinação dos pesos do método de Newton-Cotes e da expressão dos erros, a Tabela 6.2 é construída.

Tabela 6.1: Fórmulas fechadas do método de Newton-Cotes

n	N	$NC_0^{(n)}$	$NC_1^{(n)}$	$NC_2^{(n)}$	$NC_3^{(n)}$	$NC_4^{(n)}$	$NC_5^{(n)}$	$NC_6^{(n)}$	Erro da Integração
1	2	1	1						$-\frac{1}{12}h^3 f''(\xi)$
2	6	1	4	1					$-\frac{1}{90}h^5 f^{IV}(\xi)$
3	8	1	3	3	1				$-\frac{3}{80}h^5 f^{IV}(\xi)$
4	90	7	32	12	32	7			$-\frac{8}{945}h^7 f^{VI}(\xi)$
5	288	19	75	50	50	75	19		$-\frac{275}{12096}h^7 f^{VI}(\xi)$
6	840	41	216	27	272	27	216	41	$-\frac{9}{1400}h^9 f^{VIII}(\xi)$

As fórmulas fechadas na Tabela 6.2 são então construídas da seguinte forna:

$$I = \int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{i=0}^{n} NC_i^{(n)} f(x_i), \text{ sendo } x_i = a+ih \text{ e } h = \frac{(b-a)}{n} \text{ para } i = 0, 1, \cdots, n.$$

Por exemplo, a fórmula de Simpson, com $n=2$, ou seja, aproximando $f(x)$ por uma parábola, resulta:

$$I = \int_a^b f(x)dx \approx \frac{(b-a)}{6}[f(x_0)+4f(x_1)+f(x_2)] = \frac{h}{3}[f(x_0)+4f(x_1)+f(x_2)].$$

■ **Exemplo 6.1** Obter a aproximação da integral $I = \int_0^{1,2} e^x dx$, usando $n=6$ da Tabela 6.2. Para este exemplo, $a=0$, $b=1,2$, $h=(b-a)/n=0,2$ e $x_i = ih$. Portanto, da Tabela 6.2, resulta:

$$I = \int_0^{1,2} e^x dx \approx \frac{1,2}{840}\left(41e^0 + 216e^{0,2} + 27e^{0,4} + 272e^{0,6} + 27e^{0,8} + 216e^{1,0} + 41e^{1,2}\right) = 2,320116929.$$

O valor exato é $I = e^{1,2} - 1 = 2,320116923$, resultando em um $ER(\%) = -2,61 \times 10^{-7}\,\%$ e $EA = 6,055 \times 10^{-9}$. ∎

De maneira similar, para as fórmulas abertas, o sistema algébrico linear para a determinação dos pesos da quadratura, com os pontos nodais $t_i = \dfrac{i+1}{n+2}$, resulta:

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ \frac{1}{n+2} & \frac{2}{n+2} & \frac{3}{n+2} & \cdots & \frac{n+1}{n+2} \\ \left(\frac{1}{n+2}\right)^2 & \left(\frac{2}{n+2}\right)^2 & \left(\frac{3}{n+2}\right)^2 & \cdots & \left(\frac{n+1}{n+2}\right)^2 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ \left(\frac{1}{n+2}\right)^n & \left(\frac{2}{n+2}\right)^n & \left(\frac{3}{n+2}\right)^n & \cdots & \left(\frac{n+1}{n+2}\right)^n \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \vdots \\ \frac{1}{n+1} \end{pmatrix}.$$

Para o exemplo da fórmula do retângulo, com $n = 0$, o sistema linear resume-se a $\omega_0 = 1$, resultando na aproximação:

$$I = (b-a)\int_0^1 f[x(t)]dt \approx (b-a)\int_0^1 p_0(t)dt = (b-a)\int_0^1 f(x_0)dt = (b-a)f(x_0),$$

que é igual à área do retângulo de base $(b-a)$ e altura $f(x_0)$, calculada no ponto médio $t_0 = 1/2$ ou $x_0 = (a+b)/2$.

Usando o mesmo procedimento também para o cálculo do erro, constrói-se a Tabela 6.2 para as fórmulas abertas de Newton-Cotes.

Tabela 6.2: Fórmulas abertas do método de Newton-Cotes

n	N	$NC_0^{(n)}$	$NC_1^{(n)}$	$NC_2^{(n)}$	$NC_3^{(n)}$	Erro da Integração
0	1	1				$\frac{1}{3}h^3 f''(\xi)$
1	2	1	1			$\frac{3}{4}h^3 f''(\xi)$
2	3	2	-1	2		$\frac{28}{90}h^5 f^{IV}(\xi)$
3	24	11	1	1	11	$\frac{95}{144}h^5 f^{IV}(\xi)$

As fórmulas abertas na Tabela 6.2 são então construídas da seguinte forna:

$$I = \int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{i=0}^{n} NC_i^{(n)} f(x_i), \text{ sendo } x_i = a+(i+1)h \text{ e } h = \frac{(b-a)}{n+2} \text{ para } i = 0, 1, \cdots, n.$$

6.2.1 Método de Simpson em Subintervalos (Regra de Simpson Composta)

Em aplicações práticas raramente se utiliza métodos de ordem superior ao do método de Simpson, preferencialmente este método é aplicado em subintervalos em que o integrando é aproximado sucessivamente por parábolas.

Na Figura 6.1 é mostrada as aproximações parabólicas sucessivas aplicada a uma função genérica $f(x)$ (curva contínua azul) no intervalo $[1, 9]$, inicialmente $f(x)$ é aproximada por duas parábolas construída na forma: $\begin{cases} \text{Parábola 1, utiliza os pontos:} (1, f(1)),\ (3, f(3)) \text{ e } (5, f(5)) \\ \text{Parábola 2 utiliza os pontos:} (5, f(5)),\ (7, f(7)) \text{ e } (9, f(9)) \end{cases}$.

Resultando na integração numérica:

$$I_1 = \frac{5-1}{6}[f(1)+4f(3)+f(5)] + \frac{9-5}{6}[f(5)+4f(7)+f(9)]$$

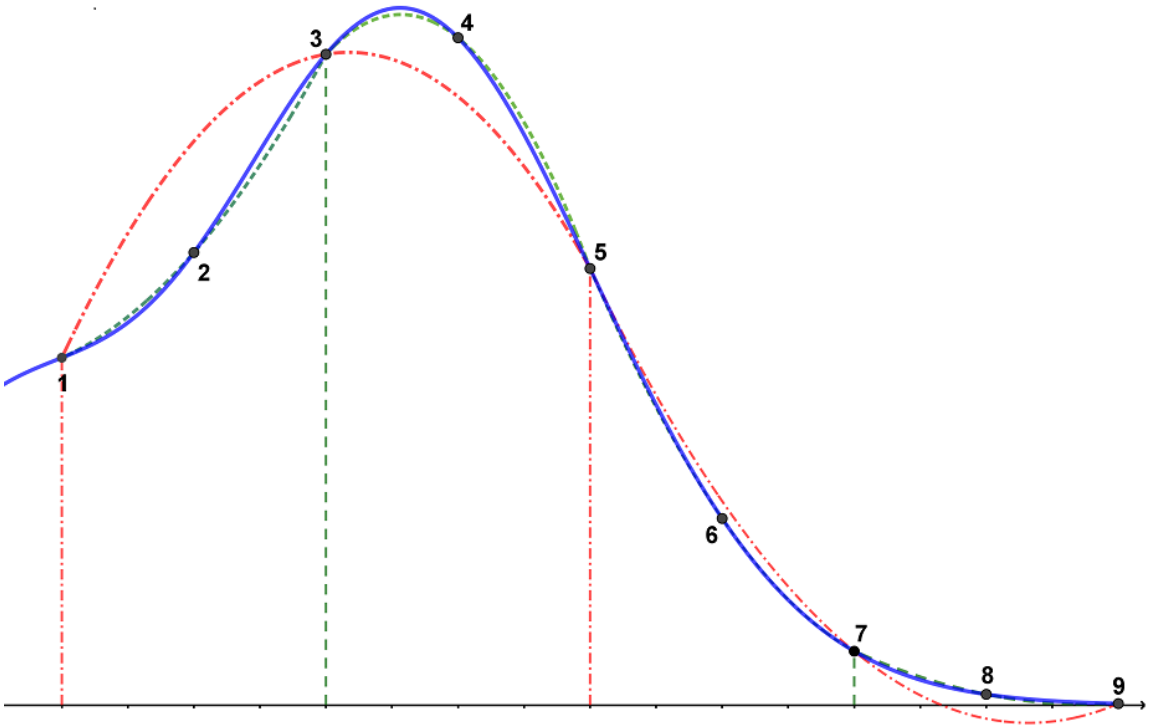


Figura 6.1: Regra de Simpson composta.

$$I_1 = 8\left[\frac{f(1)+4f(3)+2f(5)+4f(7)+f(9)}{12}\right].$$

Os valores dos módulos dos erros das duas integrações são apresentados na tabela abaixo:

$\|Erro\|_1$	$\|Erro\|_2$	$\|Erro\|_{total}$
$\frac{h_1^5}{90}f^{(IV)}(\xi_1)$	$\frac{h_1^5}{90}f^{(IV)}(\xi_2)$	$(9-1)\frac{h_1^4}{180}f^{(IV)}(\xi^*)$

Na qual $h_1 = \frac{b-a}{2N} = \frac{9-1}{4} = 2$, com $N = 2$ sendo o número de subintervalos, e o módulo do erro total foi calculado de acordo com:

$$\frac{h_1^5}{90}f^{(IV)}(\xi_1)+\frac{h_1^5}{90}f^{(IV)}(\xi_2) \leq \frac{h_1^5}{90}2f_{max}^{(IV)}(\xi) = 4\frac{h_1^4}{90}f^{(IV)}(\xi^*)$$

em que $1 \leq \xi_1 \leq 5, 5 \leq \xi_2 \leq 9$ e $1 \leq \xi^* \leq 8$, pois $2h_1 = 4 = \frac{(9-1)}{2}$ e

$$f^{(IV)}(\xi_1)+f^{(IV)}(\xi_2) \leq 2f_{max}^{(IV)}(\xi) = 2f^{(IV)}(\xi^*).$$

A seguir $f(x)$ é aproximada por quatro parábolas na forma:

$$\begin{cases} \text{Parábola 1, utiliza os pontos:}(1,f(1)),\ (2,f(2)) \text{ e } (3,f(3)) \\ \text{Parábola 2 utiliza os pontos:}(3,f(3)),\ (4,f(4)) \text{ e } (5,f(5)) \\ \text{Parábola 3 utiliza os pontos:}(5,f(5)),\ (6,f(6)) \text{ e } (7,f(7)) \\ \text{Parábola 4 utiliza os pontos:}(7,f(7)),\ (8,f(8)) \text{ e } (9,f(9)) \end{cases}.$$

Resultando na integração numérica:

$$I_2 = \frac{3-1}{6}[f(1)+4f(2)+f(3)] + \frac{5-3}{6}[f(3)+4f(4)+f(5)] + \frac{7-5}{6}[f(5)+4f(6)+f(7)] + \frac{9-7}{6}[f(7)+4f(8)+f(9)],$$

ou agrupando os termos:

$$I_2 = 8\left[\frac{f(1)+4f(2)+2f(3)+4f(4)+2f(5)+4f(6)+2f(7)+4f(8)+f(9)}{24}\right].$$

Os valores dos módulos dos erros das quatro integrações são representados na tabela abaixo:

$\|Erro\|_1$	$\|Erro\|_2$	$\|Erro\|_3$	$\|Erro\|_4$	$\|Erro\|_{total}$
$\frac{h_2^5}{90}f^{(IV)}(\xi_1)$	$\frac{h_2^5}{90}f^{(IV)}(\xi_2)$	$\frac{h_2^5}{90}f^{(IV)}(\xi_3)$	$\frac{h_2^5}{90}f^{(IV)}(\xi_4)$	$(9-1)\frac{h_2^4}{180}f^{(IV)}(\xi^*)$

O módulo do erro total foi calculado de acordo com:

$$\frac{h_2^5}{90}f^{(IV)}(\xi_1)+\frac{h_2^5}{90}f^{(IV)}(\xi_2)+\frac{h_2^5}{90}f^{(IV)}(\xi_3)+\frac{h_2^5}{90}f^{(IV)}(\xi_2) \leq \frac{h_2^5}{90}4f_{max}^{(IV)}(\xi) = 4\frac{h_2^4}{90}f^{(IV)}(\xi^*)$$

em que $1 \leq \xi_1 \leq 3 \leq \xi_2 \leq 5 \leq \xi_3 \leq 7 \leq \xi_4 \leq 9$ e $1 \leq \xi^* \leq 8$,
pois $4h_2 = 4 = \frac{(9-1)}{2}$ e $f^{(IV)}(\xi_1)+f^{(IV)}(\xi_2)+f^{(IV)}(\xi_3)+f^{(IV)}(\xi_4) \leq 4f_{max}^{(IV)}(\xi) = 4f^{(IV)}(\xi^*)$.

Desse modo, em ambos os casos o módulo do erro total da integração é: $|Erro|_{total} = Kh^4f^{(IV)}(\xi)$ sendo $h = \frac{b-a}{2N}$ a distância entre os pontos nodais empregados, N o número de parábolas empregadas, $a \leq \xi \leq b$, a o limite inferior e b o limite superior da integração.

Baseado na expressão do módulo do erro total da integração, Richardson[3] propôs o seguinte procedimento, denominado de *Extrapolação de Richardson*:

$$\begin{cases} |Erro|_{total}^{(1)} = I_{melhor} - I_1 = Kh_1^4f^{(IV)}(\xi) \\ |Erro|_{total}^{(2)} = I_{melhor} - I_2 = Kh_2^4f^{(IV)}(\xi) = \frac{Kh_1^4f^{(IV)}(\xi)}{16} \text{ com } h_2 = \frac{h_1}{2} \end{cases},$$

ou seja: $I_{melhor} - I_2 = \frac{I_{melhor} - I_1}{16} \Rightarrow I_{melhor} = \frac{16I_2 - I_1}{15}$

De acordo com estas características o método de Simpson em subintervalos pode ser implementado pelo algoritmo:

- ETAPA 0: especificação de a, b, N (número inicial de parábolas), δ (critério de convergência) e ε (menor valor do passo de integração).
- ETAPA 1: cálculo da primeira integral numérica (com N parábolas):

  $S_0 \longleftarrow f(a)+f(b)$

  $h \longleftarrow \frac{b-a}{2N}$

  $S_{impar} \longleftarrow \sum_{j=1}^{N} f[a+(2j-1)h]$

  $S_{par} = \begin{cases} \sum_{j=1}^{N-1} f[a+2jh] \text{ se } N > 1 \\ 0 \text{ se } N = 1 \end{cases}$

  $I \longleftarrow \frac{h}{3}(S_0 + 4S_{impar} + 2S_{par})$

[3] Lewis Fry Richardson (1881-1953).

- ETAPA 2: processo recursivo:

  Faça:

  $$I_{velho} \longleftarrow I$$
  $$N \longleftarrow N+N$$
  $$h \longleftarrow \frac{h}{2}$$
  $$S_{par} \longleftarrow S_{par}+S_{impar}$$
  $$S_{impar} \longleftarrow \sum_{j=1}^{N} f[a+(2j-1)h]$$
  $$I \longleftarrow \frac{h}{3}(S_0+4S_{impar}+2S_{par})$$

  Enquanto $|I-I_{velho}|>\delta$ e $|h|>\varepsilon$

- ETAPA 3: cálculo final da integral numérica (*extrapolação de Richardson*):

$$I \longleftarrow I+\frac{I-I_{velho}}{15}$$

A **Extrapolação de Richardson** empregada para melhorar o resultado do método de Simpson em subintervalos pode ser estendida para qualquer fórmula de integração aplicada em subintervalos. Se I_N e E_N são, respectivamente, a integral numérica com N subintervalos e o seu erro, então o valor exato da integral é dado por:

$$I = I_{N_1}+E_{N_1} = I_{N_2}+E_{N_2}$$

Como $E_N \propto h^m f^{(m)}(\xi) = N^{-m} f^{(m)}(\xi)$, se considerarmos que $f^{(m)}(\xi_{N_1}) = f^{(m)}(\xi_{N_2})$, então:

$$E_{N_2} = \left(\frac{N_1}{N_2}\right)^m E_{N_1} = \left(\frac{h_2}{h_1}\right)^m E_{N_1}.$$

Desse modo, pode-se obter uma boa estimativa para E_{N_1} a partir das integrais numéricas I_{N_1} e I_{N_2}:

$$E_{N_1} = \frac{I_{N_2}-I_{N_1}}{1-(N_1/N_2)^m},$$

resultando na fórmula geral da extrapolação de Richardson:

$$I_{extrapolado} = I_{N_1}+\frac{I_{N_2}-I_{N_1}}{1-(N_1/N_2)^m} = \frac{I_{N_2}-(N_1/N_2)^m I_{N_1}}{1-(N_1/N_2)^m}.$$

6.2.2 Método de Romberg

Este método de integração numérica foi desenvolvido por Romberg[4] em 1955 e consiste na associação recursiva da regra do trapézio composta com a extrapolação de Richardson.

Assim para computar numericamente a integral: $\mathbf{I} = \int_a^b f(x)dx$, inicia-se a integração numérica pela regra do trapézio usando apenas um intervalo, aumentando-se, sucessivamente, o número de subintervalos de acordo com:

[4] Werner Romberg (1909-2003).

$$\begin{cases} I^{(0)} = h^{(0)}\left(\dfrac{f(a)+f(b)}{2}\right), h^{(0)} = (b-a), x_0 = a \text{ e } x_1 = b = x_0 + h^{(0)} \\ I^{(1)} = h^{(1)}\left(\dfrac{f(a)+f(b)}{2} + f(x_1)\right), h^{(1)} = \dfrac{h^{(0)}}{2}, x_0 = a \text{ e } x_1 = x_0 + h^{(1)} \\ I^{(2)} = h^{(2)}\left(\dfrac{f(a)+f(b)}{2} + f(x_1) + f(x_2)\right), h^{(2)} = \dfrac{h^{(1)}}{2}, x_0 = a, x_k = x_0 + kh^{(2)} \\ \text{para } k = 1,2 \\ \vdots\vdots\vdots \\ I^{(N)} = h^{(N)}\left(\dfrac{f(a)+f(b)}{2} + \displaystyle\sum_{k=1}^{N} f(x_k)\right), h^{(N)} = \dfrac{h^{(N-1)}}{2}, x_0 = a, x_k = x_0 + kh^{(N)} \\ \text{para } k = 1,2,\cdots,N \end{cases}.$$

Como o método do trapézio apresenta $|Erro|_{total} = Kh^2$, a extrapolação de Richardson, para as duas primeiras integrações trapezoidais, seria: $I^{(*)} = \dfrac{4I^{(1)} - I(0)}{3} = I^{(1)} + \dfrac{I^{(1)} - I(0)}{3}$ que apresenta $|Erro|_{total} = Kh^4$, note que esta avaliação equivale ao método de Simpson pois:

$$I^{(0)} = h^{(0)}\left(\frac{f(a)+f(b)}{2}\right) \text{ e } I^{(1)} = \frac{h^{(0)}}{2}\left[\frac{f(a)+f(b)}{2} + f\left(\frac{a+b}{2}\right)\right]$$

Assim: $I^{(*)} = \dfrac{4I^{(1)} - I(0)}{3} = \dfrac{h^{(1)}}{3}\left[f(a) + 4f\left(\dfrac{a+b}{2}\right) + f(b)\right]$.

Adotando a notação: $R_{k,1} = I^{(k)}$, para $k = 0,1,2,\cdots,N$,

$R_{0,1} = h^{(0)}\left(\dfrac{f(a)+f(b)}{2}\right), h^{(0)} = (b-a), x_0 = a$ e $x_1^{(0)} = b = x_0 + h^{(0)}$.

Para: $k = 1,2,\cdots,N \longrightarrow R_{k,1} = h^{(k)}\left(\dfrac{f(a)+f(b)}{2} + \displaystyle\sum_{j=1}^{k} f(x_j^{(k)})\right), h^{(k)} = \dfrac{h^{(k-1)}}{2}$,

$x_0 = a, x_j^{(k)} = x_0 + jh^{(k)}$ para $j = 1,2,\cdots,k$.

A seguir aplica-se a extrapolação de Richardson, resultando em:

$R_{k,2} = R_{k+1,1} + \dfrac{R_{k+1,1} - R_{k,1}}{3}$ para $k = 0,1,2,\cdots,(N-1)$, que apresenta erro de quarta ordem.

Aplicando-se novamente a extrapolação de Richardson, resulta:

$R_{k,3} = R_{k+1,2} + \dfrac{R_{k+1,2} - R_{k,2}}{15}$ para $k = 0,1,2,\cdots,(N-2)$, e assim sucessivamente:

$R_{k,n+1} = R_{k+1,n} + \dfrac{R_{k+1,n} - R_{k,n}}{4^n - 1}$ para $k = 0,1,2,\cdots,(N-n)$.

Até $n = N \Rightarrow R_{0,N+1} = R_{1,N} + \dfrac{R_{1,N} - R_{0,N}}{4^N - 1}$ que é a avaliação numérica final da integral.

Conforme descrito o Método de Romberg pode ser implementado pelo algoritmo:

- ETAPA 0: especificação de a, b, e N (número total de trapézios).
- ETAPA 1: cálculo da primeira integral numérica (com 1 trapézio):

  $S_0 \longleftarrow \dfrac{f(a)+f(b)}{2}$

  $h \longleftarrow b - a$

  $S \longleftarrow 0$

  $n \longleftarrow 1$

  $m \longleftarrow N - 1$

  $R_{0,0} \longleftarrow hS_0$

- ETAPA 2: cálculo das integrais trapezoidais em subintervalos:

  Faça para $k = 1, 2, \cdots, m$
  $h \longleftarrow \dfrac{h}{2}$
  $S \longleftarrow S + \sum_{j=1}^{n} f[a + (2j-1)h]$
  $n \longleftarrow n + n$
  $R_{k,0} \longleftarrow h(S_0 + S)$

- ETAPA 3: cálculo das sucessivas extrapolações de Richardson:

  Faça para $n = 1, 2, \cdots, m$
  Faça para $k = 0, 1, \cdots, m-n$
  $$R_{k,n} \longleftarrow R_{k+1,n-1} + \frac{R_{k+1,n-1} - R_{k,n-1}}{4^n - 1}$$

- ETAPA 4: cálculo final da integral numérica:

  $I \longleftarrow R_{0,m}$.

O método de Romberg pode também ser implementado com a busca do valor de N que satisfaça a uma acurácia desejada, para isto basta aplicar o algoritmo:

- ETAPA 0: especificação de a, b, N (número inicial de trapézios) e δ (critério de convergência).
- ETAPA 1: cálculo da integral numérica pelo método de Romberg com N trapézios:
  $I \longleftarrow \text{Romberg}(f, a, b, N)$
  $flag \longleftarrow 0$
- ETAPA 2: cálculo da integral numérica pelo método de Romberg com $N+1$ trapézios

  Enquanto $flag = 0$
  $I_{velho} \longleftarrow I$
  $N \longleftarrow N + 1$
  $I \longleftarrow \text{Romberg}(f, a, b, N)$
  $flag \longleftarrow 1$ se $|I - I_{velho}| < \delta$

A integração trapezoidal em subintervalos seguinte teorema:

Teorema 6.2.1 — Teorema da Integração Trapezoidal em Subintervalos. Se $f(x)$ é uma função analítica no intervalo $(a,,)$ então a função

$$I(h) = h\left[\frac{f(a)+f(b)}{2} + \sum_{j=1}^{N_i-1} f(a+jh)\right] \text{ em que } N_i = \frac{b-a}{h},$$

admite a representação: $I(h) = I_0 + I_2h^2 + I_4h^4 + I_6h^6 + \cdots$.

A seguinte propriedade é decorrente desse teorema: $I_0 = \lim_{h \longrightarrow 0} I(h) = \int_a^b f(x)dx$.

Baseado nesse teorema e na propriedade acima, uma modificação do método de Romberg pode ser obtida pela interpolação polinomial de Lagrange na forma:

$I(h) \approx p_n(x) = \sum_{j=0}^{n} \ell_j(x) I^{(j)}$ em que $x = h^2$, $I^{(j)}$ é o valor da integração trapezoidal usando o passo $h_j = \dfrac{b-a}{2^j}$, e os pontos nodais da interpolação são: $x_j = \dfrac{h_0^2}{4^j}$, $h_0 = (b-a)$ para $j = 0, 1, 2, \cdots, n$,

resultando em:

$$I_0 = I(0) = \int_a^b f(x)dx \approx p_n(0) = \sum_{j=0}^{n} \ell_j(0)\, I^{(j)} = \sum_{j=0}^{n} \omega_j I^{(j)}.$$

Na tabela a seguir listam-se os pontos nodais com os correspondentes valores da função.

$x_0 = h_0^2$	$x_1 = \frac{h_0^2}{4}$	$x_2 = \frac{h_0^2}{4^2}$	$x_3 = \frac{h_0^2}{4^3}$	$x_4 = \frac{h_0^2}{4^4}$	$\cdots$	$x_j = \frac{h_0^2}{4^j}$
$I^{(0)}$	$I^{(1)}$	$I^{(2)}$	$I^{(3)}$	$I^{(4)}$	$\cdots$	$I^{(j)}$

Assim, os valores de ω_j são obtidos a partir de:

$$\omega_j = \ell_j(0) = \sum_{k=0\neq j}^{n} \frac{0 - x_k}{x_j - x_k} = \sum_{k=0\neq j}^{n} \frac{1}{1 - x_j/x_k} = \sum_{k=0\neq j}^{n} \frac{1}{1 - 4^{k-j}}.$$

Denominando este método como **Método de Romberg-Lagrange** que pode ser implementado pelo algoritmo:

- ETAPA 0: especificação de a, b, e N (número total de trapézios).
- ETAPA 1: cálculo da primeira integral numérica (com 1 trapézio):

  $S_0 \longleftarrow \dfrac{f(a)+f(b)}{2}$
  $h \longleftarrow b - a$
  $S \longleftarrow 0$
  $m \longleftarrow 1$
  $I^{(0)} \longleftarrow hS_0$

  Para $k = 0, 1, \cdots, N$
  $\omega_k \longleftarrow 1$
  para $j = 0, 1, 2, \cdots, N$
  se $j \neq k$, $\omega_k \longleftarrow \omega_k \left(\dfrac{1}{1 - 4^{j-k}} \right)$

- ETAPA 2: cálculo das integrais trapezoidais em subintervalos:

  Faça: para $k = 1, 2, \cdots, N$
  $h \longleftarrow \dfrac{h}{2}$
  $S \longleftarrow S + \sum_{j=1}^{m} f\,[a + (2j-1)h]$
  $m \longleftarrow m + m$
  $I^{(k)} \longleftarrow h\,(S_0 + S)$

- ETAPA 3: cálculo final da integral numérica:

$$I = \sum_{k=0}^{N} \omega_k I^{(k)}$$

O método de Romberg-Lagrange pode também ser implementado com busca do valor de N que satisfaça a um critério de precisão, para isto basta aplicar o seguinte algoritmo:

- ETAPA 0: especificação de a, b, N (número inicial de trapézios) e δ (critério de convergência).
- ETAPA 1: cálculo da integral numérica pelo método de Romberg com N trapézios:
  $I \longleftarrow$ Romberg_Lagrange(f, a, b, N)
  $flag \longleftarrow 0$
- ETAPA 2: cálculo da integral numérica pelo método de Romberg-Lagrange com $N+1$ trapézios

Enquanto $flag = 0$
$I_{velho} \longleftarrow I$
$N \longleftarrow N+1$
$I \longleftarrow$ Romberg_Lagrange(f,a,b,N)
$flag \longleftarrow 1$ se $|I - I_{velho}| < \delta$

■ **Exemplo 6.2** Comparação de Três Métodos de Integração em Subintervalos. Deseja-se confrontar o desempenho dos métodos de Simpson em subintervalos, de Romberg e de Romberg-Lagrange, utilizando o mesmo critério de acurácia $\delta = 10^{-13}$ para a integração de várias funções em diferentes intervalos. Sendo o erro listado a diferença entre a integral *exata* e a integral numérica.

(1) $I = \int_1^3 e^x dx = 17,36725509$ { Simpson: Erro= $-7,1 \times 10^{-15}$, n° de parábolas=2^{11} ; Romberg: Erro= 0, n° de trapézios=7 ; Romberg-Lagrange: Erro= 0, n° de trapézios=6

(2) $I = \int_0^2 e^{x^2} dx = 16,45262777$ { Simpson: Erro= $-7,1 \times 10^{-15}$, n° de parábolas=2^{13} ; Romberg: Erro= $-3,6 \times 10^{-15}$, n° de trapézios=10 ; Romberg-Lagrange: Erro= 0, n° de trapézios=9

(3) $I = \int_0^2 x^2 e^{x^2} dx = 46,37183615$ { Simpson: Erro= $5,7 \times 10^{-14}$, n° de parábolas=2^{14} ; Romberg: Erro= $-1,4 \times 10^{-14}$, n° de trapézios=10 ; Romberg-Lagrange: Erro= 0, n° de trapézios=9

(4) $I = \int_0^{12} \frac{\text{sen}(x)}{x} dx = 1,50497124$ { Simpson: Erro= $2,5 \times 10^{-15}$, n° de parábolas=2^{12} ; Romberg: Erro= $-1,1 \times 10^{-15}$, n° de trapézios=10 ; Romberg-Lagrange: Erro= 0, n° de trapézios=9

(5) $I = \sqrt{\frac{2}{\pi}} \int_0^5 \text{sen}(x^2) dx = 0,42121705$ { Simpson: Erro= $5,4 \times 10^{-15}$, n° de parábolas=2^{14} ; Romberg: Erro= 0, n° de trapézios=12 ; Romberg-Lagrange: Erro= 0, n° de trapézios=11

(6) $I = \frac{2}{\sqrt{(\pi)}} \int_0^2 e^{-x^2} dx = \text{erf}(2) = 0,99532227$ { Simpson: Erro= $8,3 \times 10^{-15}$, n° de parábolas=2^{10} ; Romberg: Erro= 0, n° de trapézios=9 ; Romberg-Lagrange: Erro= 0, n° de trapézios=8

(7) $I = \int_0^5 J_0(x^2) dx = 1,03298599$ { Simpson: Erro= $1,4 \times 10^{-14}$, n° de parábolas=2^{13} ; Romberg: Erro= 0, n° de trapézios=11 ; Romberg-Lagrange: Erro= 0, n° de trapézios=10

Deve-se destacar que os métodos de integração numérica em subintervalos só convergem se a função $f(x)$ for analítica em **todo** o intervalo $[a,b]$. Por exemplo, a função $f(x) = \text{sen}(\sqrt{x})$ no intervalo $[0,\pi^2]$ não é analítica em $x = 0$, pois $\frac{df(x)}{dx} = \frac{\cos(\sqrt{x})}{2\sqrt{x}}$ que assume o valor infinito em $x = 0$. Apesar dessa singularidade, essa função apresenta integral analítica e igual a:

$$I = \int_0^{\pi^2} \text{sen}(\sqrt{x}) dx = 2\left[\text{sen}(\sqrt{x}) - \sqrt{x}\cos(\sqrt{x})\right]_0^{\pi^2} = -2\pi = -6,28318531.$$

Essa singularidade pode ser removida pela mudança da variável x para $x = u^2$, assim $\text{sen}(\sqrt{x})dx = 2u\,\text{sen}(u)du$ e

$$I = \int_0^{\pi^2} \text{sen}(\sqrt{x}) dx = 2\int_0^{\pi} u\,\text{sen}(u) du = 2\left[\text{sen}(u) - u\cos(u)\right]_0^{\pi} = -2\pi = -6,28318531.$$

O novo integrando: $g(u) = 2u\,\text{sen}(u)$ é agora analítico em todo o intervalo, podendo ser integrado

numericamente pelos métodos anteriores, resultando em:

$$\begin{cases} \text{Simpson: Erro= } 3,6\cdot 10^{-15},\ n^{\circ}\text{ de parábolas=}2^{12} \\ \text{Romberg: Erro= } -2,7\cdot 10^{-15},\ n^{\circ}\text{ de trapézios=8} \\ \text{Romberg-Lagrange: Erro= } 0,\ n^{\circ}\text{ de trapézios=7} \end{cases}.$$

■

6.3 Método de Quadratura de Gauss

De forma distinta aos métodos de integração de Newton-Cotes, em que se empregam pontos nodais igualmente espaçados, os métodos de quadratura tipo Gauss empregam pontos *livres* no interior do intervalo de integração que assegurem a maior acurácia possível. De uma forma geral o método da quadratura de Gauss é expresso por:

$$I=\int_{-1}^{+1} f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i),$$

em que ω_i são os pesos da quadratura e x_i as respectivas abscissas, sendo $\omega_i > 0$ e $-1 < x_i < +1$.

No Capítulo 3 foi visto que a interpolação de Hermite, quando se empregam n pontos nodais: $0 < x_1 < x_2 < \cdots < x_n < 1$, é expressa por:

$$p_{2n-1}(x)=\sum_{j=1}^{n} \ell_j(x)^2 f(x_j) + \left[\sum_{j=1}^{n} \frac{f'(x_j)-2A_{j,j}f(x_j)}{p'_{nodal}(x_j)}\ell_j(x)\right] p_{nodal}(x),$$

em que: $p_{nodal}(x)=\prod_{k=1}^{n}(x-x_k)$ é um polinômio da grau n, $\ell_j(x)=\prod_{k=1\neq j}^{n}\frac{x-x_k}{x_j-x_k}$ é um polinômio de grau $(n-1)$ e $A_{j,j}=\left.\frac{d\ell_j(x)}{dx}\right|_{x=x_j}$. Esta interpolação pode também ser expressa na forma: $p_{2n-1}(x)=\sum_{j=1}^{n}\ell_j(x)^2 f(x_j)+q_{n-1}(x)p_{nodal}(x)$, sendo $q_{n-1}(x)$ um polinômio de grau $(n-1)$. O erro desta interpolação é expresso por:

$$R_n(x)=p^2_{nodal}(x)\frac{1}{(2n)!}\left.\frac{d^{2n}f(t)}{dt^{2n}}\right|_{t=\xi(x)},$$

permitindo expressar: $f(x)=\sum_{j=1}^{n}\ell_j(x)^2 f(x_j)+q_{n-1}(x)p_{nodal}(x)+p^2_{nodal}(x)\frac{1}{(2n)!}\left.\frac{d^{2n}f(t)}{dt^{2n}}\right|_{t=\xi(x)}$.

Assim:

$$I=\int_{-1}^{+1} f(x)dx=\sum_{i=1}^{n}\omega_i f(x_i)+\int_0^1 q_{n-1}(x)p_{nodal}(x)dx+\frac{1}{(2n)!}\int_{-1}^{+1} p^2_{nodal}(x)\left.\frac{d^{2n}f(t)}{dt^{2n}}\right|_{t=\xi(x)}dx,$$

sendo: $\omega_i=\int_{-1}^{+1}\ell_i(x)^2dx>0$.

Pelo teorema do valor médio tem-se:

$$\int_{-1}^{+1} p^2_{nodal}(x)\left.\frac{d^{2n}f(t)}{dt^{2n}}\right|_{t=\xi(x)}dx=\left.\frac{d^{2n}f(t)}{dt^{2n}}\right|_{t=\bar{\xi}(x)}\int_{-1}^{+1}p^2_{nodal}(x)dx.$$

Selecionando as abscissas ou pontos nodais tais que:

$$\int_{-1}^{+1} x^k p_{nodal}(x)dx \text{ para } k=0,1,2,\cdots,(n-1) \Rightarrow \int_{-1}^{+1} q_{n-1}(x)p_{nodal}(x)dx=0.$$

O polinômio que apresenta esta propriedade é o polinômio de Legendre, que pode ser gerado recursivamente por:

$$P_i(x) = \frac{(2i-1)xP_{i-1}(x) - (i-1)P_{i-2}(x)}{i} \text{ para } i = 2, \cdots, (n-1) \text{ com } P_0(x) = 1 \text{ e } P_1(x) = x,$$

$i = 2 \Rightarrow P_2(x) = \dfrac{3x^2-1}{2}$, além disto: $\displaystyle\int_{-1}^{+1} P_i^2(x)dx = \frac{2}{2i+1}$.

Entretanto, o $p_{nodal}(x) = \displaystyle\prod_{k=1}^{n}(x-x_k) = x^n + \cdots$ é um polinômio de grau n cujo coeficiente de x^n é igual à unidade, assim $p_{nodal}(x) = \dfrac{P_n(x)}{c_n^{(n)}}$ em que $c_n^{(n)}$ é o coeficiente de x^n no polinômio de Legendre de grau n que é igual a: $c_n^{(n)} = \dfrac{(2n)!}{2^n(n!)^2}$, implicando em:

$$\int_{-1}^{+1} p_{nodal}^2(x)dx = \frac{2}{(2i+1)\left(c_n^{(n)}\right)^2} = \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^2}.$$

Resultando finalmente em:

$$I = \sum_{i=1}^{n} \omega_i f(x_i) + R_n \text{ com } \omega_i = \int_{-1}^{+1} \ell_i(x)^2 dx = \frac{2}{(1-x_i^2)\left[P_n'(x_i)\right]^2} = \frac{2(1-x_i^2)}{\left[nP_{n-1}(x_i)\right]^2}$$

e $R_n = \dfrac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi)$ com $\xi \in (-1, +1)$.

Para ilustrar o procedimento o valor de $n = 2$ é considerado, as abscissas são as raízes de

$$P_2(x) = \frac{3x^2-1}{2} = 0 \longrightarrow \begin{cases} x_1 = -\dfrac{1}{\sqrt{3}} \\ x_2 = +\dfrac{1}{\sqrt{3}} \end{cases} \quad \text{assim : } \begin{cases} \ell_1(x) = \dfrac{x_2 - x}{x_2 - x_1} = \dfrac{1}{2}\left(1 - \sqrt{3}x\right) \\ \ell_2(x) = \dfrac{x - x_1}{x_2 - x_1} = \dfrac{1}{2}\left(1 + \sqrt{3}x\right) \end{cases}.$$

Os pesos da quadratura são: $\begin{cases} \omega_1 = \displaystyle\int_{-1}^{+1} \ell_1^2(x)dx = 1 \\ \omega_2 = \displaystyle\int_{-1}^{+1} \ell_2^2(x)dx = 1 \end{cases}$.

Dando origem a: $I = \displaystyle\int_{-1}^{+1} f(x)dx \approx f(x_1) + f(x_2)$ que é exata para funções polinomiais em x até terceiro grau.

■ **Exemplo 6.3** Como ilustração da quadratura de Gauss com duas abscissas, considere a função $f(x) = 42x^3 - 57x^2 - 36x + 20$, assim: $I = \displaystyle\int_{-1}^{+1} f(x)dx = 2 = f(x_1) + f(x_2)$. Como $f(x_1) = 1 + \dfrac{22\sqrt{3}}{3}$ e $f(x_2) = 1 - \dfrac{22\sqrt{3}}{3}$, confirma-se que $f(x_1) + f(x_2) = 2$.

O polinômio interpolador de primeiro grau é:

$$p_1(x) = \frac{1}{2}\left(1 - \sqrt{3}x\right) f(x_1) + \frac{1}{2}\left(1 + \sqrt{3}x\right) f(x_2) = \frac{f(x_1) + f(x_2)}{2} + \sqrt{3}\left(\frac{f(x_2) - f(x_1)}{2}\right) x.$$

Substituindo os valores de $f(x_1)$ e $f(x_2)$, resulta em $p_1(x) = 1 - 22x$ conforme representado na Figura 6.2.

■

Pelo fato de as abscissas deste método serem as raízes do polinômio de Legendre, o método é denominado de **Método de Quadratura de Gauss-Legendre** e alguns valores das abscissas e dos pesos são apresentados a seguir.

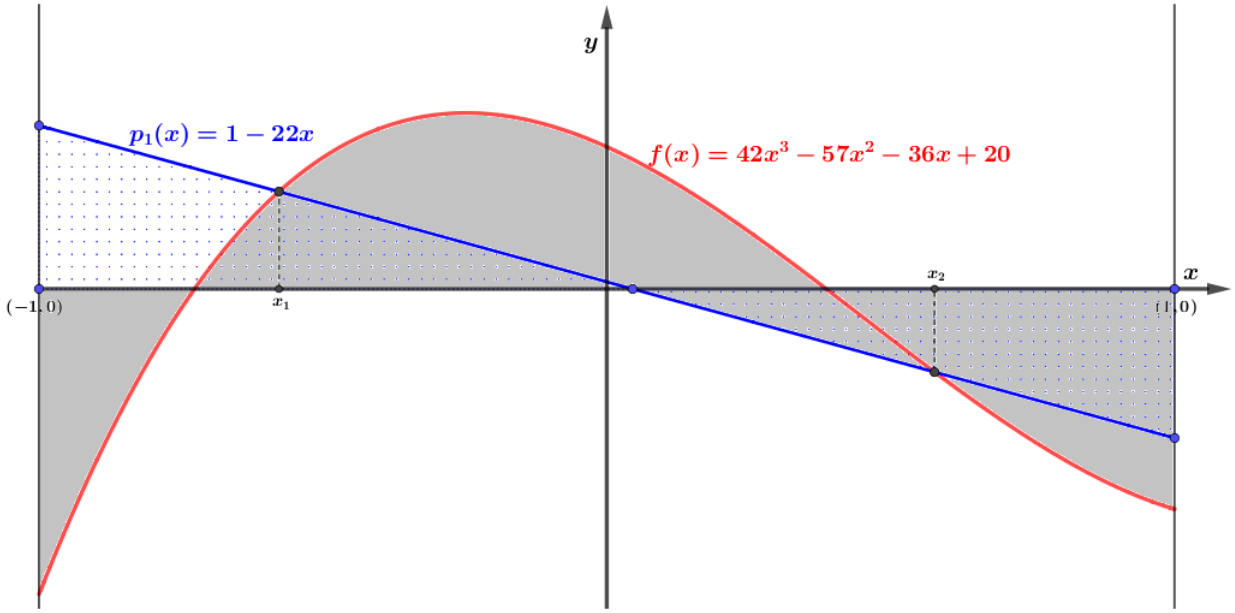


Figura 6.2: Quadratura de Gauss de um polinômio de terceiro grau.

$$\bullet\ n = 2, \mathbf{x} = \begin{pmatrix} -\frac{1}{\sqrt{3}} \\ +\frac{1}{\sqrt{3}} \end{pmatrix} \text{ e } \omega = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; \qquad \bullet\ n = 3, \mathbf{x} = \begin{pmatrix} -\sqrt{\frac{3}{5}} \\ 0 \\ +\sqrt{\frac{3}{5}} \end{pmatrix} \text{ e } \omega = \frac{1}{9}\begin{pmatrix} 5 \\ 8 \\ 5 \end{pmatrix};$$

$$\bullet\ n = 4, \mathbf{x} = \begin{pmatrix} -\frac{1}{7}\sqrt{21+14\sqrt{\frac{6}{5}}} \\ -\frac{1}{7}\sqrt{21-14\sqrt{\frac{6}{5}}} \\ +\frac{1}{7}\sqrt{21-14\sqrt{\frac{6}{5}}} \\ +\frac{1}{7}\sqrt{21+14\sqrt{\frac{6}{5}}} \end{pmatrix} \text{ e } \omega = \frac{1}{36}\begin{pmatrix} 18-\sqrt{30} \\ 18+\sqrt{30} \\ 18+\sqrt{30} \\ 18-\sqrt{30} \end{pmatrix};$$

$$\bullet\ n = 5, \mathbf{x} = \begin{pmatrix} -\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}} \\ -\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}} \\ 0 \\ +\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}} \\ +\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}} \end{pmatrix} \text{ e } \omega = \frac{1}{900}\begin{pmatrix} 322-13\sqrt{70} \\ 322+13\sqrt{70} \\ 512 \\ 322+13\sqrt{70} \\ 322-13\sqrt{70} \end{pmatrix}.$$

A inclusão de uma das extremidades ou de ambas as extremidades do intervalo da origem a formas modificadas do método de quadratura de Gauss, descritas a seguir:

(1) **Métodos de Quadratura de Gauss-Radau**[5] - inclusão de uma das extremidades.

(a) Inclusão da extremidade inferior: $x=-1$.

Dando origem à expressão: $I=\dfrac{2}{(n+1)^2}f(-1)+\sum_{i=1}^{n}\omega_i f(x_i)+R_n$, as abscissas internas são as raízes do polinômio de grau n descrito por:

$Q_n(x)=\dfrac{P_{n+1}(x)+P_n(x)}{1+x}$ os correspondentes pesos são:

$\omega_i=\dfrac{1}{(1-x_i)[P_n'(x_i)]^2}=\dfrac{1-x_i}{[(n+1)P_n(x_i)]^2}$ e

$R_n=\dfrac{2^{2n+1}(n+1)(n!)^4}{[(2n+1)!]^3}f^{(2n+1)}(\xi)$ com $\xi\in(-1,+1)$.

As abscissas e os pesos das duas primeiras quadraturas de Gauss-Radau com inclusão de $x=-1$ são:

$n=1\Rightarrow \mathbf{x}=\begin{pmatrix}-1\\ \dfrac{1}{3}\end{pmatrix}$ e $\omega=\begin{pmatrix}1/2\\ 3/2\end{pmatrix}$;

$n=2\Rightarrow \mathbf{x}=\begin{pmatrix}-1\\ \dfrac{1-\sqrt{6}}{5}\\ \dfrac{1+\sqrt{6}}{5}\end{pmatrix}$ e $\omega=\dfrac{1}{18}\begin{pmatrix}4\\ 16+\sqrt{6}\\ 16-\sqrt{6}\end{pmatrix}$.

(b) Inclusão da extremidade superior: $x=+1$.

Dando origem à expressão: $I=\sum_{i=1}^{n}\omega_i f(x_i)+\dfrac{2}{(n+1)^2}f(+1)+R_n$, as abscissas internas são as raízes do polinômio de grau n descrito por:

$Q_n(x)=\dfrac{P_{n+1}(x)-P_n(x)}{1-x}$ os correspondentes pesos são:

$\omega_i=\dfrac{1}{(1+x_i)[P_n'(x_i)]^2}=\dfrac{1+x_i}{[(n+1)P_n(x_i)]^2}$ e

$R_n=\dfrac{2^{2n+1}(n+1)(n!)^4}{[(2n+1)!]^3}f^{(2n+1)}(\xi)$ com $\xi\in(-1,+1)$.

As abscissas e os pesos das duas primeiras Quadraturas de Gauss-Radau com inclusão de $x=-1$ são:

$n=1\Rightarrow \mathbf{x}=\begin{pmatrix}-1/3\\ +1\end{pmatrix}$ e $\omega=\begin{pmatrix}3/2\\ 1/2\end{pmatrix}$;

$n=2\Rightarrow \mathbf{x}=\begin{pmatrix}\dfrac{-1-\sqrt{6}}{5}\\ \dfrac{-1+\sqrt{6}}{5}\\ +1\end{pmatrix}$ e $\omega=\dfrac{1}{18}\begin{pmatrix}16+\sqrt{6}\\ 16-\sqrt{6}\\ 4\end{pmatrix}$.

(2) **Métodos de Quadratura de Gauss-Lobatto**[6] - Inclusão de ambas as extremidades: $x=-1$ e $x=+1$.

Dando origem à expressão:

$I=\dfrac{2}{(n+1)(n+2)}f(-1)+\sum_{i=1}^{n}\omega_i f(x_i)+\dfrac{2}{(n+1)(n+2)}f(+1)+R_n$, as abscissas internas são as raízes do polinômio de grau n resultante da diferenciação do polinômio de Legendre de

[5] Jean Charles Rodolphe Radau (1835-1911).

[6] Rehuel Lobatto (1797-1866).

grau $(n+1)$: $Q_n(x) = \dfrac{dP_{n+1}(x)}{dx}$, os correspondentes pesos são: $\omega_i = \dfrac{2}{(n+1)(n+2)[P_{n+1}(x_i)]^2} = \dfrac{2(n+1)}{(n+2)[P'_n(x_i)]^2}$ e

$R_n = \dfrac{2^{2n+3}(n+1)(n+2)(n!)^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi)$ com $\xi \in (-1, +1)$.

As abscissas e os pesos das três primeiras Quadraturas de Gauss-Lobatto são:

$n = 1 \Rightarrow \mathbf{x} = \begin{pmatrix} -1 \\ 0 \\ +1 \end{pmatrix}$ e $\omega = \dfrac{1}{3}\begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix}$ (procedimento análogo ao do Método de Simpson);

$n = 2 \Rightarrow \mathbf{x} = \begin{pmatrix} -1 \\ -\dfrac{1}{\sqrt{5}} \\ +\dfrac{1}{\sqrt{5}} \\ +1 \end{pmatrix}$ e $\omega = \dfrac{1}{6}\begin{pmatrix} 1 \\ 5 \\ 5 \\ 1 \end{pmatrix}$;

$n = 3 \Rightarrow \mathbf{x} = \begin{pmatrix} -1 \\ -\sqrt{\dfrac{3}{7}} \\ 0 \\ +\sqrt{\dfrac{3}{7}} \\ +1 \end{pmatrix}$ e $\omega = \dfrac{1}{90}\begin{pmatrix} 9 \\ 49 \\ 64 \\ 49 \\ 9 \end{pmatrix}$.

■ **Exemplo 6.4** Os pesos e abscissas das quadraturas tipo Gauss podem também ser determinadas por procedimentos semelhantes aos de determinação dos pesos dos métodos de integração tipo Newton-Cotes, neste exemplo são mostrados tais procedimentos que dispensam o conhecimento prévio dos polinômios ortogonais e o emprego da interpolação de Lagrange.

1. **Métodos de Quadratura de Gauss**

$$\int_{-1}^{+1} x^k dx = \frac{1-(-1)^{k+1}}{k+1} = \sum_{i=1}^{n} \omega_i x_i^k, \text{ para } k = 0, 1, 2, \cdots, (2n-1)$$

(a) $n = 1 : \begin{cases} \omega_1 = 2 \\ \omega_1 x_1 = 0 \Rightarrow x_1 = 0 \end{cases}$;

(b) $n = 2 : \begin{cases} \omega_1 + \omega_2 = 2 \\ \omega_1 x_1 + \omega_2 x_2 = 0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \\ \omega_1 x_1^3 + \omega_2 x_2^3 = 0 \end{cases}$

Considerando o polinômio de 2° grau: $p_2(x) = (x-x_1)(x-x_2) = x^2 + c_1 x + c_0 \Rightarrow p_2(x_1) = p_2(x_2) = 0$.

$$(+)\begin{cases} \omega_1 + \omega_2 = 2 \longleftarrow \times c_0 \\ \omega_1 x_1 + \omega_2 x_2 = 0 \longleftarrow \times c_1 \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \end{cases} \quad \Big| \quad (+)\begin{cases} \omega_1 x_1 + \omega_2 x_2 = 0 \longleftarrow \times c_0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \longleftarrow \times c_1 \\ \omega_1 x_1^3 + \omega_2 x_2^3 = 0 \end{cases}$$

$$\omega_1 p_2(x_1) + \omega_2 p_2(x_2) = 0 = 2c_0 + \frac{2}{3} \quad \Big| \quad \omega_1 x_1 p_2(x_1) + \omega_2 x_2 p_2(x_2) = 0 = \frac{2}{3} c_1$$

Assim $c_0 = -\dfrac{1}{3}$ e $c_1 = 0 \Rightarrow p_2(x) = x^2 - \dfrac{1}{3}$, logo $x_1 = -\dfrac{1}{\sqrt{3}}$ e $x_2 = \dfrac{1}{\sqrt{3}}$. Obtidas as

abscissas, os pesos são então determinados resolvendo o sistema:

$$\begin{cases} \omega_1 + \omega_2 = 2 \\ \omega_1 x_1 + \omega_2 x_2 = (\omega_1 - \omega_2) x_1 = 0 \end{cases} \Rightarrow \omega_1 = \omega_2 = 1$$

(c) $n = 3$:
$$\begin{cases} \omega_1 + \omega_2 + \omega_3 = 2 \\ \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \\ \omega_1 x_1^5 + \omega_2 x_2^5 + \omega_3 x_3^5 = 0 \end{cases}$$

Considerando o polinômio de 3° grau: $p_3(x) = (x-x_1)(x-x_2)(x-x_3) = x^3 + c_2 x^2 + c_1 x + c_0 \Rightarrow p_3(x_1) = p_3(x_2) = p_3(x_3) = 0$.

$$(+)\begin{cases} \omega_1 + \omega_2 + \omega_3 = 2 \longleftarrow \times c_0 \\ \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_1 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_2 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \end{cases}$$

$$\omega_1 p_3(x_1) + \omega_2 p_3(x_2) + \omega_3 p_3(x_3) = 0 = 2c_0 + \frac{2}{3}c_2$$

$$(+)\begin{cases} \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_1 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \longleftarrow \times c_2 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \end{cases}$$

$$\omega_1 x_1 p_3(x_1) + \omega_2 x_2 p_3(x_2) + \omega_3 x_3 p_3(x_3) = 0 = \frac{2}{3}c_1 + \frac{2}{5} \Rightarrow c_1 = -\frac{3}{5}$$

$$(+)\begin{cases} \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_0 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \longleftarrow \times c_1 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \longleftarrow \times c_2 \\ \omega_1 x_1^5 + \omega_2 x_2^5 + \omega_3 x_3^5 = 0 \end{cases}$$

$$\omega_1 x_1^2 p_3(x_1) + \omega_2 x_2^2 p_3(x_2) + \omega_3 x_3^2 p_3(x_3) = 0 = \frac{2}{3}c_0 + \frac{2}{5}c_2$$

$$\begin{cases} 2c_0 + \dfrac{2}{3}c_2 = 0 \\ \dfrac{2}{3}c_0 + \dfrac{2}{5}c_2 = 0 \end{cases} \Rightarrow c_0 = c_2 = 0 \text{ como } c_1 = -\frac{3}{5} \Rightarrow p_3(x) = x^3 - \frac{3}{5}x = x\left(x^2 - \frac{3}{5}\right).$$

Assim $x_1 = -\sqrt{\dfrac{3}{5}}$, $x_2 = 0$ e $x_2 = \sqrt{\dfrac{3}{5}}$. Para as abscissas, resolve-se o sistema linear:

$$\begin{cases} \omega_1 + \omega_2 + \omega_3 = 2 \\ (\omega_1 - \omega_3) x_1 = 0 \\ (\omega_1 + \omega_3) x_1^2 = (\omega_1 + \omega_3)\dfrac{3}{5} = \dfrac{2}{3} \end{cases} \Rightarrow \omega_1 = \omega_3 = \frac{5}{9} \text{ e } \omega_2 = \frac{8}{9}$$

2. **Métodos de Quadratura de Gauss-Radau com** $x_0 = -1$.

$$\int_{-1}^{+1} x^k dx = \frac{1-(-1)^{k+1}}{k+1} = \sum_{i=0}^{n} \omega_i x_i^k, \text{ para } k = 0, 1, 2, \cdots, 2n$$

(a) $n = 1$:
$$\begin{cases} \omega_0 + \omega_1 = 2 \\ \omega_0 x_0 + \omega_1 x_1 = 0 \\ \omega_0 x_0^2 + \omega_1 x_1^2 = \dfrac{2}{3} \end{cases}$$

Considerando o polinômio de 2° grau: $p_2(x) = (x-x_0)(x-x_1) = x^2 + c_1x + c_0 \Rightarrow$ $p_2(x_1) = 0$ e $p_2(x_0) = p_2(-1) = 0 = 1 - c_1 + c_0$.

$$(+)\begin{cases} \omega_0 + \omega_1 = 2 \longleftarrow \times c_0 \\ \omega_0 x_0 + \omega_1 x_1 = 0 \longleftarrow \times c_1 \\ \omega_0 x_0^2 + \omega_1 x_1^2 = \dfrac{2}{3} \end{cases}$$

$$\omega_0 p_2(x_0) + \omega_1 p_2(x_1) = 0 = 2c_0 + \frac{2}{3}$$

Logo $c_0 = -\dfrac{1}{3}$ mas $c_1 = 1 + c_0 = \dfrac{2}{3} \Rightarrow p_2(x) = x^2 + \dfrac{2}{3}x - \dfrac{1}{3} = (x+1)(x-1/3) \Rightarrow$

$x_1 = \dfrac{1}{3}$ e $\begin{cases} \omega_0 + \omega_1 = 2 \\ -\omega_0 + \dfrac{\omega_1}{3} = 0 \end{cases} \Rightarrow \begin{cases} \omega_0 = 1/2 \\ \omega_1 = 3/2 \end{cases}$.

(b) $n = 2$: $\begin{cases} \omega_0 + \omega_1 + \omega_2 = 2 \\ \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 = 0 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 = 0 \\ \omega_0 x_0^4 + \omega_1 x_1^4 + \omega_2 x_2^4 = \dfrac{2}{5} \end{cases}$

Considerando o polinômio de 3° grau: $p_3(x) = (x-x_0)(x-x_1)(x-x_2) = x^3 + c_2x^2 +$ $c_1x + c_0 \Rightarrow p_3(x_1) = p_3(x_2) = 0$ e $p_3(x_0) = p_3(-1) = -1 + c_2 - c_1 + c_0 = 0$.

$$(+)\begin{cases} \omega_0 + \omega_1 + \omega_2 = 2 \longleftarrow \times c_0 \\ \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 = 0 \longleftarrow \times c_1 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \longleftarrow \times c_2 \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 = 0 \end{cases} \qquad (+)\begin{cases} \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 = 0 \longleftarrow \times c_0 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \longleftarrow \times c_1 \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 = 0 \longleftarrow \times c_2 \\ \omega_0 x_0^4 + \omega_1 x_1^4 + \omega_2 x_2^4 = \dfrac{2}{5} \end{cases}$$

$$\omega_0 p_3(x_0) + \omega_1 p_3(x_1) + \omega_2 p_3(x_2) = 0 = 2c_0 + \frac{2}{3}c_2$$

$$\omega_0 x_0 p_3(x_0) + \omega_1 x_1 p_3(x_1) + \omega_2 x_2 p_3(x_2) = 0 = \frac{2}{3}c_1 + \frac{2}{5} \Rightarrow c_1 = -\frac{3}{5}$$

$$\begin{cases} 2c_0 + \dfrac{2}{3}c_2 = 0 \\ c_0 + c_2 = 1 + c_1 = 1 - \dfrac{3}{5} = \dfrac{2}{5} \end{cases} \Rightarrow \begin{cases} c_0 = -\dfrac{1}{5} \\ c_2 = \dfrac{3}{5} \end{cases} .$$

$p_3(x) = x^3 + \dfrac{3}{5}x^2 - \dfrac{3}{5}x - \dfrac{1}{5} = (x+1)\left(x^2 - \dfrac{2}{5}x - \dfrac{1}{5}\right) \Rightarrow \begin{cases} x_1 = \dfrac{1-\sqrt{6}}{5} \\ x_2 = \dfrac{1+\sqrt{6}}{5} \end{cases}$ e com o

sistema linear:

$$\begin{cases} \omega_0 + \omega_1 + \omega_2 = 2 \\ -\omega_0 + \omega_1\left(\dfrac{1-\sqrt{6}}{5}\right) + \omega_2\left(\dfrac{1+\sqrt{6}}{5}\right) = 0 \\ \omega_0 + \omega_1\left(\dfrac{1-\sqrt{6}}{5}\right)^2 + \omega_2\left(\dfrac{1+\sqrt{6}}{5}\right)^2 = \dfrac{2}{3} \end{cases} \Rightarrow \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{pmatrix} = \frac{1}{18}\begin{pmatrix} 4 \\ 16+\sqrt{6} \\ 16-\sqrt{6} \end{pmatrix}.$$

3. **Métodos de Quadratura de Gauss-Radau com** $x_{n+1} = +1$.

$$\int_{-1}^{+1} x^k dx = \frac{1-(-1)^{k+1}}{k+1} = \sum_{i=1}^{n+1} \omega_i x_i^k, \text{ para } k = 0, 1, 2, \cdots, 2n$$

(a) $n = 1$: $\begin{cases} \omega_1 + \omega_2 = 2 \\ \omega_1 x_1 + \omega_2 x_2 = 0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \end{cases}$

Considerando o polinômio de 2° grau: $p_2(x) = (x-x_1)(x-x_2) = x^2 + c_1 x + c_0 \Rightarrow$ $p_2(x_1) = 0$ e $p_2(x_2) = p_2(+1) = 0 = 1 + c_1 + c_0$.

$$(+)\begin{cases} \omega_1 + \omega_2 = 2 \longleftarrow \times c_0 \\ \omega_1 x_1 + \omega_2 x_2 = 0 \longleftarrow \times c_1 \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3} \end{cases}$$

$$\omega_1 p_2(x_1) + \omega_2 p_2(x_2) = 0 = 2c_0 + \frac{2}{3}$$

Logo $c_0 = -\dfrac{1}{3}$ mas $c_1 = -1 - c_0 = -\dfrac{2}{3} \Rightarrow p_2(x) = x^2 - \dfrac{2}{3}x - \dfrac{1}{3} = (x-1)(x+1/3) \Rightarrow$

$x_1 = -\dfrac{1}{3}$ e $\begin{cases} \omega_1 + \omega_2 = 2 \\ -\dfrac{\omega_1}{3} + \omega_2 = 0 \end{cases} \Rightarrow \begin{cases} \omega_1 = 3/2 \\ \omega_2 = 1/2 \end{cases}$.

(b) $n = 2$: $\begin{cases} \omega_1 + \omega_2 + \omega_3 = 2 \\ \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \end{cases}$

Considerando o polinômio de 3° grau: $p_3(x) = (x-x_1)(x-x_2)(x-x_3) = x^3 + c_2 x^2 +$ $c_1 x + c_0 \Rightarrow p_3(x_1) = p_3(x_2) = 0$ e $p_3(x_3) = p_3(+1) = 1 + c_2 + c_1 + c_0 = 0$.

$$(+)\begin{cases} \omega_1 + \omega_2 + \omega_3 = 2 \longleftarrow \times c_0 \\ \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_1 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_2 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \end{cases}$$

$$\omega_1 p_3(x_1) + \omega_2 p_3(x_2) + \omega_3 p_3(x_3) = 0 = 2c_0 + \frac{2}{3}c_2$$

$$(+)\begin{cases} \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_0 \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_1 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \longleftarrow \times c_2 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \end{cases}$$

$$\omega_1 x_1 p_3(x_1) + \omega_2 x_2 p_3(x_2) + \omega_3 x_3 p_3(x_3) = 0 = \frac{2}{3}c_1 + \frac{2}{5} \Rightarrow c_1 = -\frac{3}{5}$$

$$\begin{cases} 2c_0 + \dfrac{2}{3}c_2 = 0 \\ c_0 + c_2 = -1 - c_1 = 1 + \dfrac{3}{5} = -\dfrac{2}{5} \end{cases} \Rightarrow \begin{cases} c_0 = \dfrac{1}{5} \\ c_2 = -\dfrac{3}{5} \end{cases} .$$

$$p_3(x) = x^3 - \frac{3}{5}x^2 - \frac{3}{5}x + \frac{1}{5} = (x-1)\left(x^2 + \frac{2}{5}x - \frac{1}{5}\right) \Rightarrow \begin{cases} x_1 = -\left(\dfrac{\sqrt{6}+1}{5}\right) \\ x_2 = \dfrac{\sqrt{6}-1}{5} \end{cases} \quad \text{e}$$

com o sistema linear:

$$\begin{cases} \omega_1+\omega_2+\omega_3=2 \\ -\omega_1\left(\dfrac{\sqrt{6}+1}{5}\right)+\omega_2\left(\dfrac{\sqrt{6}-1}{5}\right)+\omega_3=0 \\ \omega_1\left(\dfrac{\sqrt{6}+1}{5}\right)^2+\omega_2\left(\dfrac{\sqrt{6}-1}{5}\right)^2+\omega_3=\dfrac{2}{3} \end{cases} \Rightarrow \begin{pmatrix}\omega_0\\ \omega_1\\ \omega_2\end{pmatrix}=\frac{1}{18}\begin{pmatrix}16-\sqrt{6}\\ 16+\sqrt{6}\\ 4\end{pmatrix}.$$

• As abscissas e os pesos das quadraturas de Gauss-Radau com a inclusão da extremidade inferior $(x_0=-1)$ e com a inclusão da extremidade superior $(x_{n+1}=+1)$ apresentam uma certa simetria, que pode ser expressa por:

Quadratura de Gauss-Radau com $x_0=-1$	Quadratura de Gauss-Radau com $x_{n+1}=+1$
Abscissas $\mathbf{x}=\begin{pmatrix}x_0=-1\\ x_1\\ x_2\\ \vdots\\ x_{n-1}\\ x_n\end{pmatrix}=\begin{pmatrix}\rho_0\\ \rho_1\\ \rho_2\\ \vdots\\ \rho_{n-1}\\ \rho_n\end{pmatrix}$	Abscissas $\mathbf{x}=\begin{pmatrix}x_1\\ x_2\\ x_3\\ \vdots\\ x_n\\ x_{n+1}=+1\end{pmatrix}=-\begin{pmatrix}\rho_n\\ \rho_{n-1}\\ \rho_{n-2}\\ \vdots\\ \rho_1\\ \rho_0\end{pmatrix}$
Pesos $\omega=\begin{pmatrix}\omega_0\\ \omega_1\\ \omega_2\\ \vdots\\ \omega_{n-1}\\ \omega_n\end{pmatrix}=\begin{pmatrix}\sigma_0\\ \sigma_1\\ \sigma_2\\ \vdots\\ \sigma_{n-1}\\ \sigma_n\end{pmatrix}$	Pesos $\omega=\begin{pmatrix}\omega_1\\ \omega_2\\ \omega_3\\ \vdots\\ \omega_n\\ \omega_{n+1}\end{pmatrix}=\begin{pmatrix}\sigma_n\\ \sigma_{n-1}\\ \sigma_{n-2}\\ \vdots\\ \sigma_1\\ \sigma_0\end{pmatrix}$

4. **Métodos de Quadratura de Gauss-Lobatto** - Inclusão de ambas as extremidades: $x_0=-1$ e $x_{n+1}=+1$.

$$\int_{-1}^{+1}x^k dx=\frac{1-(-1)^{k+1}}{k+1}=\sum_{i=0}^{n+1}\omega_i x_i^k,\ \text{para } k=0,1,2,\cdots,(2n+1)$$

(a) $n=1:\begin{cases}\omega_0+\omega_1+\omega_2=2\\ \omega_0x_0+\omega_1x_1+\omega_2x_2=0\\ \omega_0x_0^2+\omega_1x_1^2+\omega_2x_2^2=\dfrac{2}{3}\\ \omega_0x_0^3+\omega_1x_1^3+\omega_2x_2^3=0\end{cases}$

Considerando o polinômio de 3° grau:
$p_3(x)=(x-x_0)(x-x_1)(x-x_2)=x^3+c_2x^2+c_1x+c_0\Rightarrow p_3(x_1)=0,$
$p_3(x_0)=p_3(-1)=-1+c_2-c_1+c_0=0$ e $p_3(x_2)=p_3(+1)=1+c_2+c_1+c_0=0.$

$$\begin{cases}c_2-c_1+c_0=(c_0+c_2)-c_1=1\\ c_2+c_1+c_0=(c_0+c_2)+c_1=-1\end{cases}\Rightarrow\begin{cases}c_0+c_2=0\\ c_1=-1\end{cases}$$

$$(+)\begin{cases}\omega_0+\omega_1+\omega_2=2 \longleftarrow \times c_0\\ \omega_0x_0+\omega_1x_1+\omega_2x_2=0 \longleftarrow \times c_1\\ \omega_0x_0^2+\omega_1x_1^2+\omega_2x_2^2=\dfrac{2}{3} \longleftarrow \times c_2\\ \omega_0x_0^3+\omega_1x_1^3+\omega_2x_2^3=0\end{cases}$$

$$\omega_0p_3(x_0)+\omega_1p_3(x_1)+\omega_2p_3(x_2)=0=2c_0+\frac{2}{3}c_2$$

$$\begin{cases} c_0 + c_2 = 0 \\ 2c_0 + \dfrac{2}{3}c_2 = 0 \end{cases} \Rightarrow c_0 = c_2 = 0.$$

Resultando em $p_3(x) = x^3 - x = x(x^2-1) \Rightarrow x_1 = 0$ e com isso o sistema linear:

$$\begin{cases} \omega_0 + \omega_1 + \omega_2 = 2 \\ -\omega_0 + \omega_2 = 0 \\ \omega_0 + \omega_2 = \dfrac{2}{3} \end{cases} \Rightarrow \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix}.$$

(b) $n = 2:$
$$\begin{cases} \omega_0 + \omega_1 + \omega_2 + \omega_3 = 2 \\ \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \\ \omega_0 x_0^4 + \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \\ \omega_0 x_0^5 + \omega_1 x_1^5 + \omega_2 x_2^5 + \omega_3 x_3^5 = 0 \end{cases}$$

Considerando o polinômio de 4° grau:
$p_4(x) = (x-x_0)(x-x_1)(x-x_2)(x-x_3) = x^4 + c_3x^3 + c_2x^2 + c_1x + c_0 \Rightarrow p_4(x_1) = p_4(x_2) = 0$, $p_4(x_0) = p_4(-1) = 1 - c_3 + c_2 - c_1 + c_0 = 0$ e $p_4(x_3) = p_4(+1) = 1 + c_3 + c_2 + c_1 + c_0 = 0$.

$$\begin{cases} -c_3 + c_2 - c_1 + c_0 = (c_0 + c_2) - (c_1 + c_3) = -1 \\ c_3 + c_2 + c_1 + c_0 = (c_0 + c_2) + (c_1 + c_3) = -1 \end{cases} \Rightarrow \begin{cases} c_0 + c_2 = -1 \\ c_1 + c_3 = 0 \end{cases}$$

$$(+)\begin{cases} \omega_0 + \omega_1 + \omega_2 + \omega_3 = 2 \longleftarrow \times c_0 \\ \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_1 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_2 \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \longleftarrow \times c_3 \\ \omega_0 x_0^4 + \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \end{cases}$$

$$\omega_0 p_4(x_0) + \omega_1 p_4(x_1) + \omega_2 p_4(x_2) + \omega_3 p_4(x_3) = 0$$
$$2c_0 + \frac{2}{3}c_2 = -\frac{2}{5}$$

$$(+)\begin{cases} \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0 \longleftarrow \times c_0 \\ \omega_0 x_0^2 + \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3} \longleftarrow \times c_1 \\ \omega_0 x_0^3 + \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0 \longleftarrow \times c_2 \\ \omega_0 x_0^4 + \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5} \longleftarrow \times c_3 \\ \omega_0 x_0^5 + \omega_1 x_1^5 + \omega_2 x_2^5 + \omega_3 x_3^5 = 0 \end{cases}$$

$$\omega_0 x_0 p_4(x_0) + \omega_1 x_1 p_4(x_1) + \omega_2 x_2 p_4(x_2) + \omega_3 x_3 p_4(x_3) = 0 \Rightarrow \frac{2}{3}c_1 + \frac{2}{5}c_3 = 0$$

$$\begin{cases} 2c_0 + \dfrac{2}{3}c_2 = 0 \\ c_0 + c_2 = -1 \end{cases} \Rightarrow \begin{pmatrix} c_0 \\ c_2 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} +1 \\ -6 \end{pmatrix} \text{ e } \begin{cases} \dfrac{2}{3}c_1 + \dfrac{2}{5}c_3 = 0 \\ c_1 + c_3 = 0 \end{cases} \Rightarrow c_1 = c_3 = 0$$

Assim $p_4(x) = x^4 - \dfrac{6}{5}x^2 + \dfrac{1}{5} = (x^2-1)\left(x^2 - \dfrac{1}{5}\right) \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \dfrac{1}{\sqrt{5}}\begin{pmatrix} -1 \\ +1 \end{pmatrix}$ e com o sistema linear:

$$\begin{cases} \omega_0 + \omega_1 + \omega_2 + \omega_3 = 2 \\ -\omega_0 - \dfrac{\omega_1}{\sqrt{5}} + \dfrac{\omega_2}{\sqrt{5}} + \omega_3 = 0 \\ \omega_0 + \dfrac{\omega_1}{5} + \dfrac{\omega_2}{5} + \omega_3 = \dfrac{2}{3} \\ -\omega_0 - \dfrac{\omega_1}{5\sqrt{5}} + \dfrac{\omega_2}{5\sqrt{5}} + \omega_3 = 0 \end{cases} \Rightarrow \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 1 \\ 5 \\ 5 \\ 1 \end{pmatrix}$$

■

6.3.1 Outras Formas de Quadratura

As diferentes formas da quadratura de Gauss apresentadas são casos particulares de um procedimento mais geral expresso por:

$$I = \int_a^b w(x)f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i),$$

sendo $w(x) > 0$ uma função contínua de $x \in (a,b)$, tal função é chamada de **função peso** que depende da natureza do problema.

Por exemplo, o valor médio de uma função distribuída uniformemente em uma esfera de raio R, $f(r/R)$ com $0 \leq r \leq R$, é calculado por:

$V_{total}\bar{f} = \left(\frac{4}{3}\pi R^3\right)\bar{f} = 4\pi\int_0^R r^2 f(r/R)dr \Rightarrow \bar{f} = 3\int_0^1 x^2 f(x)dx$ em que $x = \frac{r}{R}$,

permitindo identificar: $w(x) = 3x^2 > 0$ e $\int_0^1 w(x)dx = 1$.

Para calcular numericamente a integral $I = \int_a^b w(x)f(x)dx$ com a maior acurácia possível, deve-se adotar as abscissas da quadratura como sendo as raízes do n-ésimo polinômio ortogonal da família: $I = \int_a^b w(x)P_i^{(w)}(x)P_j^{(w)}(x)dx = \mathbf{C}_i^{(w)}\delta_{i,j} > 0$. A geração de cada família de polinômios ortogonais será função apenas da função peso (identificada pelo sobrescrito (w)) e do intervalo de integração. Os polinômios ortogonais mais empregados em métodos de quadratura são apresentados na Tabela 6.3.

Tabela 6.3: Polinômios ortogonais em relação à função peso $w(x)$ e o intervalo de integração (a,b).

a	b	$w(x)$	**Polinômio**
-1	+1	1	Legendre: $P_n(x)$
-1	+1	$(1-x)^\alpha(1+x)^\beta$	Jacobi: $P_n^{(\alpha,\beta)}(x)$
-1	+1	$\frac{1}{\sqrt{1-x^2}}$	Chebyshev (1° tipo): $T_n(x)$
-1	+1	$\sqrt{1-x^2}$	Chebyshev (2° tipo): $U_n(x)$
0	$+\infty$	e^{-x}	Laguerre: $L_n(x)$
$-\infty$	$+\infty$	e^{-x^2}	Hermite: $H_n(x)$

Propriedades específicas de cada uma dessas famílias de polinômios são encontradas em Manuais Matemáticos e em diversos domínios da Internet.

6.4 Métodos Numéricos para Cômputo de Integrais Duplas

Para calcular numericamente a integral dupla, também chamada de **Cubatura numérica**, $I = \int_a^b \int_{c(x)}^{d(x)} f(x,y)dydx$. considera-se preliminarmente a função $G(x) = \int_{c(x)}^{d(x)} f(x,y)dy$ e, a seguir, $I = \int_a^b G(x)dx$.

A discretização de ambas variáveis de integração é feita de acordo com: $x_i = a + i\left(\frac{b-a}{n}\right) = a + ih_x$ para $i = 0, 1, 2\cdots, n$, sendo $h_x = \frac{b-a}{n}$, e $y_j^{(i)} = c(x_i) + j\left(\frac{d(x_i)-c(x_i)}{m}\right) = c(x_i) + jh_x^{(i)}$ para $j = 0, 1, 2\cdots, m$, sendo $h_y^{(i)} = \frac{d(x_i)-c(x_i)}{m}$.

Após a construção da *malha* de discretização os métodos de integração, descritos anteriormente, são aplicados nos subintervalos em cada direção. As formas algorítmicas de cada um desses métodos são apresentadas a seguir.

6.4.1 Regra de Simpson Composta para Cômputo de Integrais Duplas

Construção da sub-rotina: $Simpson_{composto}[f(x),a,b,c(x),d(x),n,m]$.

- ETAPA 0: especificação de $f(x)$, a, b, $c(x)$, $d(x)$, n e m (número de parábolas na direção x e y, respectivamente).
- ETAPA 1: cálculos preliminares:
  $n \longleftarrow n+n$
  $m \longleftarrow m+m$
  $h_x \longleftarrow \dfrac{b-a}{n}$
  $S \longleftarrow 0$
  $S_{impar} \longleftarrow 0$
  $S_{par} \longleftarrow 0$
- ETAPA 2: cálculo das áreas das parábolas nas duas direções:

  Para $i = 0, 1, \cdots, n,$ faça
  $x \longleftarrow a + i\,h_x$
  $h_y \longleftarrow \dfrac{d(x)-c(x)}{m}$
  $T \longleftarrow f[x,c(x)] + f[x,d(x)]$
  $T_{impar} \longleftarrow 0$
  $T_{par} \longleftarrow 0$
  $I \longleftarrow \dfrac{h}{3}(S_0 + 4S_{impar} + 2S_{par})$
  Para $j = 1, \cdots, m-1,$ faça
  $y \longleftarrow c(x) + j\,h_y$
  $F \longleftarrow f(x,y)$
  $\begin{cases} T_{par} \longleftarrow T_{par} + F \text{ se } j \text{ for par;} \\ T_{impar} \longleftarrow T_{impar} + F \text{ se } j \text{ for ímpar;} \end{cases}$
  $G_y \longleftarrow \dfrac{h_y}{3}(T + 2T_{par} + 4T_{impar})$
  Se $i = 0$ ou $i = n,$ faça
  $S \longleftarrow S + G_y$
  Senão $\begin{cases} S_{par} \longleftarrow S_{par} + G_y \text{ se } i \text{ for par;} \\ S_{impar} \longleftarrow S_{impar} + G_y \text{ se } i \text{ for ímpar;} \end{cases}$
- ETAPA 3: cálculo final da integral numérica:

  $I \longleftarrow \dfrac{h_x}{3}(S + 2S_{par} + 4S_{impar})$

A aplicação recursiva do procedimento, visando a atender um critério de convergência, é feita pelo algoritmo.

- ETAPA 0: especificação de $f(x)$, a, b, $c(x)$, $d(x)$, n, m (número inicial de parábolas na direção x e y, respectivamente) e δ (critério de convergência).
- ETAPA 1: cálculo preliminar:
  $I \longleftarrow Simpson_{composto}[f(x),a,b,c(x),d(x),n,m]$
  $Flag \longleftarrow 0$
- ETAPA 2: cálculo do valor numérico da integral com a duplicação dos pontos nodais:

Enquanto $Flag = 0$, faça
$n \longleftarrow n+n$
$m \longleftarrow m+m$
$I_{novo} \longleftarrow Simpson_{composto}[f(x),a,b,c(x),d(x),n,m]$
$\Delta \longleftarrow I_{novo} - I$
$Flag \longleftarrow 1$ se $|\Delta| < \delta$
$I \longleftarrow I_{novo}$

- ETAPA 3: cálculo final da integral numérica:

$I \longleftarrow I + \dfrac{\Delta}{15}$ (Extrapolação de Richardson)

6.4.2 Regra de Romberg Composta para Cômputo de Integrais Duplas

Antes de apresentar as duas formas do método de Romberg para integrais duplas, será apresentado o procedimento de construção de trapézios, nas duas direções, o que compõe a sub-rotina: $Trapezio_{composto}[f(x),a,b,c(x),d(x),n]$.

- ETAPA 0: especificação de $f(x)$, a, b, $c(x)$, $d(x)$ e n (número de trapézios nas direções x e y).
- ETAPA 1: cálculos preliminares:
  $h_x \longleftarrow \dfrac{b-a}{n}$
  $S \longleftarrow 0$
  $S_0 \longleftarrow 0$
- ETAPA 2: cálculo das áreas dos trapézios nas duas direções:

Para $i = 0, 1, \cdots, n,$ faça
$x \longleftarrow a + i h_x$
$h_y \longleftarrow \dfrac{d(x)-c(x)}{n}$
$T_0 \longleftarrow f[x,c(x)] + f[x,d(x)]$
$T \longleftarrow 0$
Para $j = 1, \cdots, n-1,$ faça
$y \longleftarrow c(x) + j h_y$
$T \longleftarrow T + f(x,y)$
$G_y \longleftarrow \dfrac{h_y}{2}(T_0 + 2T)$
Se $i = 0$ ou $i = n,$ faça
$S_0 \longleftarrow S_0 + G_y$
Senão $S \longleftarrow S + G_y$

- ETAPA 3: cálculo final da integral numérica:

$I \longleftarrow \dfrac{h_x}{2}(S_0 + 2S)$

Algoritmo para a Regra de Romberg para Cômputo de Integrais Duplas

Compondo a sub-rotina: $Romberg_{composto}[f(x),a,b,c(x),d(x),n]$.

- ETAPA 0: especificação de $f(x)$, a, b, $c(x)$, $d(x)$ e n (número de vezes que se reduz o passo à metade).
- ETAPA 1: cálculo preliminar:
  $m \longleftarrow 1$
  $I_0 \longleftarrow Trapezio_{composto}[f(x),a,b,c(x),d(x),m]$

- ETAPA 2: cálculo do valor numérico da integral com a duplicação dos pontos nodais:

  Para $k = 1, 2, \cdots, n,$ faça
  $m \longleftarrow m + m$
  $I_k \longleftarrow Trapezio_{composto}[f(x), a, b, c(x), d(x), m]$

  Para $k = 0, \cdots, n-1,$ faça
  $I_k \longleftarrow I_{k+1} + \dfrac{I_{k+1} - I_k}{4^n - 1}$ (Extrapolação de Richardson)

- ETAPA 3: cálculo final da integral numérica:

  $I_{Romberg} \longleftarrow I_0$.

A aplicação recursiva do procedimento, visando a atender um critério de convergência, é feita pelo algoritmo.

- ETAPA 0: especificação de $f(x), a, b, c(x), d(x)$ e n (número inicial de vezes que se reduz o passo à metade) e δ (critério de convergência).
- ETAPA 1: cálculo preliminar:
  $I \longleftarrow Romberg_{composto}[f(x), a, b, c(x), d(x), n]$
  $Flag \longleftarrow 0$
- ETAPA 2: cálculo do valor numérico da integral com o aumento de n:

  Enquanto $Flag = 0,$ faça
  $n \longleftarrow n + 1$
  $I_{novo} \longleftarrow Romberg_{composto}[f(x), a, b, c(x), d(x), n]$
  $Flag \longleftarrow 1$ se $|I_{novo} - I| < \delta$
  $I \longleftarrow I_{novo}$

- ETAPA 3: cálculo final da integral numérica:

  $I_{Romberg} \longleftarrow I$.

Algoritmo para a Regra de Romberg-Lagrange para Cômputo de Integrais Duplas

Compondo a sub-rotina: Romberg_Lagrange$_{composto}[f(x), a, b, c(x), d(x), n]$.

- ETAPA 0: especificação de $f(x), a, b, c(x), d(x)$ e n (número de vezes que se reduz o passo à metade).
- ETAPA 1: cálculos preliminares:
  $m \longleftarrow 1$

  Para $i = 0, 1, \cdots, n,$ faça
  $\omega_i \longleftarrow 1$
  Para $j = 0, 1, \cdots, n,$ faça
  se $j \neq i,$ $\omega_i \longleftarrow \dfrac{\omega_i}{1 - 4^{j-i}}$

  $I_0 \longleftarrow Trapezio_{composto}[f(x), a, b, c(x), d(x), m]$

- ETAPA 2: cálculo do valor numérico da integral com a duplicação dos pontos nodais:

  Para $k = 1, 2, \cdots, n,$ faça
  $m \longleftarrow m + m$
  $I_k \longleftarrow Trapezio_{composto}[f(x), a, b, c(x), d(x), m]$

- ETAPA 3: cálculo final da integral numérica:

  $$I_{\text{Romberg\_Lagrange}} \longleftarrow \sum_{k=0}^{n} \omega_k I_k$$

A aplicação recursiva do procedimento, visando a atender um critério de convergência, é feita pelo algoritmo.

- ETAPA 0: especificação de $f(x), a, b, c(x), d(x)$ e n (número inicial de vezes que se reduz o passo à metade) e δ (critério de convergência).
- ETAPA 1: cálculo preliminar:
  $I \longleftarrow \text{Romberg\_Lagrange}_{composto}[f(x), a, b, c(x), d(x), n]$
  $Flag \longleftarrow 0$
- ETAPA 2: cálculo do valor numérico da integral com o aumento de n:

> Enquanto $Flag = 0$, faça
> $n \longleftarrow n+1$
> $I_{novo} \longleftarrow \text{Romberg\_Lagrange}_{composto}[f(x), a, b, c(x), d(x), n]$
> $Flag \longleftarrow 1$ se $|I_{novo} - I| < \delta$
> $I \longleftarrow I_{novo}$

- ETAPA 3: cálculo final da integral numérica:

  $I_{\text{Romberg\_Lagrange}} \longleftarrow I$.

6.4.3 Método da Quadratura de Gauss para Cômputo de Integrais Duplas

Considerando a integração $I = \int_a^b \int_{c(x)}^{d(x)} f(x,y)dydx$, define-se:

$$G(x) = \int_{c(x)}^{d(x)} f(x,y)dy = \frac{d(x)-c(x)}{2} \int_{z=-1}^{z=+1} f[x, y(z)]dz$$

Em que: $y(z) = \dfrac{[d(x)-c(x)]z + [d(x)+c(x)]}{2}$. Essa integral é calculada numericamente pelo método de quadratura de Gauss, resultando em:

$G_{Gauss} = \dfrac{d(x)-c(x)}{2} \sum_{i=1}^{n} \omega_i^{(n)} f[x, y(z_i)]$, sendo $\omega_i^{(n)}$ os pesos da quadratura e z_i as abscissas, que são as raízes do polinômio de Legendre de grau n.

Calculando a seguir a integral:

$$I = \int_a^b G(x)dx \approx \int_a^b G_{Gauss}(x)dx \approx \frac{b-a}{2} \sum_{j=1}^{m} \omega_j^{(m)} G_{Gauss}(x_j), \text{ sendo } x_j = \frac{(b-a)\xi_j + (b+a)}{2},$$

$\omega_j^{(m)}$ os pesos da quadratura e ξ_j as abscissas, que são as raízes do polinômio de Legendre de grau m. Resultando:

$$I_{Gauss} = \frac{b-a}{2} \sum_{j=1}^{m} \omega_j^{(m)} \left[\frac{d(x_j)-c(x_j)}{2} \sum_{i=1}^{n} \omega_i^{(n)} f[x_j, y(z_i)] \right].$$

6.5 Cômputo de Integrais com Singularidades

Integrais singulares, assim denominadas no lugar de impróprias, são aquelas em que o integrando apresenta um comportamento singular em uma ou em ambas as extremidades do intervalo de integração, ou integrais em que pelo menos um dos limites da integração é infinito. Na literatura há inúmeros métodos para tratar esses problemas, entretanto, no presente texto, os tratamentos dos mesmos são feitos casuisticamente e demonstrados em exemplos específicos. Em todos esses exemplos, os resultados obtido pelos três métodos aplicados no Exemplo 6.2, com o critério de convergência $\delta = 10^{-13}$, foram os mesmos reportados a seguir.

(1) **Integrais com singularidades nas extremidades**

(a) Neste primeiro exemplo, o integrando não é limitado no limite inferior da integração, como na integral:
$I = \int_a^b \frac{f(x)}{(x-a)^p} dx$, em que $0 \leq p < 1$ e $f(x)$ é uma função analítica em a.
Esta singularidade pode ser removida pela troca da variável de integração x por:
$du = (1-p)\frac{dx}{(x-a)^p} \Rightarrow u(x) = (x-a)^{1-p} = \begin{cases} 0 \text{ se } x = a \\ u^* = (b-a)^{1-p} \text{ se } x = b \end{cases}$
e $x(u) = a + u^{\frac{1}{1-p}}$. A integral, em termos da nova variável de integração, transforma-se em:
$I = \frac{1}{1-p} \int_0^{u^*} f[x(u)]du$, que não apresenta singularidade nas extremidades e pode ser integrada numericamente pelos procedimentos usuais. Caso a singularidade deste mesmo tipo ocorrer em $x = b$, procedimento análogo é empregado.
Para ilustrar o procedimento a integral $I = \int_0^1 \frac{e^x}{\sqrt{x}} dx$ é considerada,
assim: $du = \frac{dx}{2\sqrt{x}} \Rightarrow u(x) = \sqrt{x} = \begin{cases} 0 \text{ se } x = 0 \\ 1 \text{ se } x = 1 \end{cases}$ e $x(u) = u^2$. A integral, em termos da nova variável de integração, transforma-se em: $I = 2\int_0^1 e^{u^2} du$ e o novo integrando, $2e^{u^2}$, não apresenta singularidade em todo domínio real. O valor numérico da integral é $2,925303491814$.

(b) Integrais envolvendo a função logarítmica em intervalos no qual o argumento é nulo, como no exemplo: $I = \int_0^1 \frac{\ln(x)}{1+x} dx$. Aplicando novamente a mudança de variável $u(x) = \sqrt{x}$:
$I = 2\int_0^1 \frac{\sqrt{x}\ln(x)}{1+x} \frac{dx}{2\sqrt{x}} = 4\int_0^1 \frac{u\ln(u)}{1+u^2} du$, que apresenta uma singularidade em $u = 0$, entretanto $\lim_{u \longrightarrow 0}[u\ln(u)] = 0$ assim, o integrando: $f(u) = \begin{cases} 0 \text{ se } u = 0 \\ 4\frac{u\ln(u)}{1+u^2} \text{ se } u \neq 0 \end{cases}$.
O valor numérico de $I = \int_0^1 f(u)du$ é igual a $-0,822467033424 = -\frac{\pi^2}{12}$ (Spiegel e Liu, 1999).
Outro exemplo, também reportado por Spiegel e Liu (1999), é $I = \int_0^1 \frac{\ln(x)}{1-x} dx = -\frac{\pi^2}{6}$.
Essa integral apresenta singularidades em ambas as extremidades. Aplicando novamente a mudança da variável de integração: $I = 2\int_0^1 \frac{\sqrt{x}\ln(x)}{1-x} \frac{dx}{2\sqrt{x}} = 4\int_0^1 \frac{u\ln(u)}{(1+u)(1-u)} du$, que apresenta singularidades em $u = 0$ e $u = 1$.
Entretanto $\lim_{u \longrightarrow 0}[u\ln(u)] = 0$ e $\lim_{u \longrightarrow 1} \frac{\ln(u)}{1-u} = -1$ (aplicando L'Hôpital[7]), resultando em:
$f(u) = \begin{cases} 0 \text{ se } u = 0 \\ -2 \text{ se } u = 1 \\ \frac{4u\ln(u)}{1-u^2} \text{ se } 0 \leq u \leq 1 \end{cases}$.

[7] Guillaume François Antoine Marquis de L'Hôpital (1661-1704).

O valor exato de $I=\int_0^1 f(u)du$ é igual a $-1,644934066848=-\frac{\pi^2}{6}$.

(2) **Integrais envolvendo uma ou ambas extremidades infinitas**

Quando a integral envolve limite superior infinito: $I=\int_a^\infty f(x)dx$, aplica-se a mudança da variável $t=\frac{1}{x+1-a}\Rightarrow t=\begin{cases}1 \text{ se } x=a\\ 0 \text{ se } x\longrightarrow\infty\end{cases}, x(t)=\frac{1+(a-1)t}{t}$ e $dx=-\frac{1}{t^2}$, resultando na integral $I=\int_0^1 \frac{f[x(t)]}{t^2}dt$, que apresenta singularidade em $t=0$, podendo ser tratada como no primeiro caso.

(a) Exemplo: calcular a integral $I=\int_1^\infty x^{-3/2}\text{sen}(1/x)dx$.

Aplicando a mudança de variável $t=\frac{1}{x}\Rightarrow I=\int_0^1 \frac{\text{sen}(t)}{\sqrt{t}}dt$, considerando agora $u=\sqrt{t}$ tem-se $t=u^2$ e $du=\frac{dt}{2\sqrt{t}}$, resultando em $I=2\int_0^1 \text{sen}(u^2)du$. O valor exato de $I=2\int_0^1 \text{sen}(u^2)du=\sqrt{2\pi}\,Si\left(\sqrt{\frac{2}{\pi}}\right)$ é igual a $0,620536603447$. Em que $Si(x)=\int_0^x \text{sen}\left(\frac{\pi t^2}{2}\right)dt$ é a integral seno de Fresnel[8].

(b) Exemplo: calcular a integral $I=\int_1^\infty \frac{1}{1+x^4}dx$.

Aplicando a mudança de variável $t=\frac{1}{1+x}\Rightarrow t=\begin{cases}1 \text{ se } x=0\\ 0 \text{ se } x\longrightarrow\infty\end{cases}$, $x(t)=\frac{1-t}{t}$ e $dx=-\frac{1}{t^2}$, resultando na integral $I=\int_0^1 \frac{t^2}{t^4+(1-t)^4}dt$.
O valor exato desta integral, $1,1107207345396$, foi obtido por integração analítica após fatoração.

(c) Exemplo (Spiegel e Liu, 1999): calcular a integral $I=\int_0^\infty e^{-x}\ln(x)dx=-\gamma$ (Constante de Euler[9] $=-0,577215664902$). Esta integral é resolvida em duas integrações sucessivas:
$I=\int_0^1 e^{-x}\ln(x)dx+\int_1^\infty e^{-x}\ln(x)dx$.
$I_1=\int_0^1 e^{-x}\ln(x)dx=2\int_0^1 e^{-x}\ln(x)\sqrt{x}\frac{dx}{2\sqrt{x}}=4\int_0^1 e^{-u^2}u\ln(u)du=\int_0^1 g(u)du$ em que $g(u)=\begin{cases}0 \text{ se } u=0\\ 4e^{-u^2}u\ln(u) \text{ se } u\neq 0\end{cases}$. O valor numérico da integral é $-0,796599599297$.
Para $I_2=\int_1^\infty e^{-x}\ln(x)dx$ a integração é feita por partes considerando:
$du=e^{-x}dx\Rightarrow u(x)=-e^{-x}$ e $v(x)=\ln(x)\Rightarrow dv(x)=\frac{dx}{x}$, como $u(x)v(x)|_1^\infty=0$ resulta
$I_2=\int_1^\infty \frac{e^{-x}}{x}dx=\int_0^1 \frac{e^{-1/t}}{t}dt=\int_0^1 f(t)dt$, em que $f(t)=\begin{cases}0 \text{ se } t=0\\ \frac{e^{-1/t}}{t} \text{ se } t\neq 0\end{cases}$.
O valor numérico desta integral é $0,219383934396$.
Então: $I=I_1+I_2=-0,577215664902=-\gamma$.

[8] Augustin Jean Fresnel (1788-1827).
[9] Leonhard Euler (1707-1783).

6.6 Problemas Propostos

Problema 6.1 O fluxo, $q(\lambda,T)d\lambda$, com que a energia radiante é emitida da superfície de um corpo negro com comprimento de onda entre λ e $\lambda+d\lambda$ é dada pela equação de Planck[10]:

$$q(\lambda,T)d\lambda = \frac{2\pi hc^2}{\lambda^5\left[\exp\left(\frac{hc}{k\lambda T}\right)-1\right]}d\lambda$$

Sendo:
$$\begin{cases} c: \text{ velocidade da luz} = 2,997925\times10^{10}\dfrac{cm}{s}; \\ h: \text{ constante de Planck} = 6,6256\times10^{-27} erg\cdot s; \\ T: \text{ temperatura } K; \\ \lambda: \text{ comprimento de onda } cm. \end{cases}$$

Calcule o fluxo total da energia emitida, em $\dfrac{erg}{cm^2 s}$, de um corpo negro entre os comprimentos de onda $\lambda_1 = 3933,666 \mathring{A}ngstr\ddot{o}m$ e $\lambda_2 = 5895,923 \mathring{A}ngstr\ddot{o}m$ às temperaturas de 2000 e 6000 K.

Problema 6.2 Em um trocador de calor de casco e tubo, vapor saturado é alimentado ao casco visando aquecer uma corrente de um fluido que escoa no tubo, de acordo com a Figura 6.3.

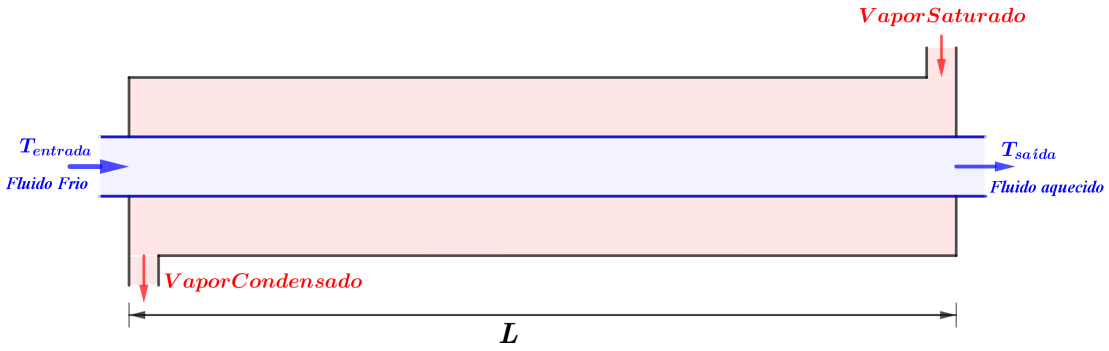


Figura 6.3: Trocador de calor de casco e tubo.

O comprimento do trocador é obtido através da integração do balanço de energia do sistema dando origem a:

$$L = \frac{W}{\pi D}\int_{T_{entrada}}^{T_{saída}}\left[\frac{c_P(T)}{h(T)(T_{vapor}-T)}\right]dT$$

Sendo:
$$\begin{cases} L: \text{ comprimento do trocador;} \\ W: \text{ vazão mássica do fluido do tubo;} \\ D: \text{ diâmetro do tubo;} \\ T: \text{ temperatura ;} \\ c_P: \text{ calor específico do fluido do tubo;} \\ h: \text{coeficiente de transferência de calor entre o tubo e o casco.} \end{cases}$$

O coeficiente h é dado através da correlação empírica:

$$h(T) = \frac{0,023k(T)}{D}\left(\frac{4W}{\pi D\mu(T)}\right)^{0,8}\left(\frac{\mu(T)c_P(T)}{k(T)}\right)^{0,4}.$$

[10] Max Karl Ernst Ludwig Planck (1858-1947).

Sendo: $\begin{cases} k: \text{ coeficiente de condutividade térmica do fluido do casco;} \\ \mu: \text{ viscosidade do fluido do casco.} \end{cases}$

Calcular o comprimento do trocador para os casos tabelados a seguir:

	Caso A	Caso B
Fluido	CO_2 em fase gasosa	Etileno glicol líquido
$W(lb_m/h)$	22,5	45000
$T_{entrada}(^\circ F)$	60	0
$T_{saída}(^\circ F)$	280 e 500	90 e 180
$T_{vapor}(^\circ F)$	550	250
D(polegadas)	0,495	1,032
$c_P(BTU/lb_m/^\circ F)$	$0,251+3,46\times10^{-5}T-\frac{14400}{(T+460)^2}$	$0,53+0,00065T$
$k(BTU/h/ft/^\circ F)$	0,0085 (32°F); 0,01815 (392°F); 0,0133 (212°F); 0,02228 (572°F)	0,153 (constante)
$\mu(lb_m/ft/h)$	$0,0332\left(\frac{T+460}{460}\right)^{0,935}$	242 (0°F); 82,1 (50°F); 30,5 (100°F); 12,6 (150°F); 5,57 (200°F)

Problema 6.3 A tabela abaixo apresenta a taxa de geração de CO_2 e a taxa de consumo de oxigênio em vários tempos, durante o processo de fermentação de *Penicillium chrysogenum*.

Tempo (h)	Taxa de Geração de CO_2 (g/h)	Taxa de Consumo de O_2 (g/h)
0	15,72	15,49
1	15,53	16,16
2	15,19	15,35
3	16,56	15,13
4	16,21	14,20
5	17,39	14,23
6	17,36	14,29
7	17,42	12,74
8	17,60	14,74
9	17,75	13,68
10	18,95	14,51

Os valores totais de CO_2 e de O_2 são determinados pela integração no tempo dessas taxas. Uma variável importante do processo, que permite avaliar a atividade metabólica do micro-organismo utilizado, é o chamado coeficiente respiratório que é a razão entre o CO_2 total formado e o oxigênio total consumido, baseado nesta informação e nos valores tabelados calcule o valor deste coeficiente ao cabo de 10 horas do processo.

Problema 6.4 O escoamento de líquidos newtonianos em tubos cilíndricos apresenta um perfil parabólico de velocidades, neste caso é comum o cômputo de integrais da forma:

$f_{medio} = 4\int_0^1 (1-r^2)rf(r)dr$, tal integral pode ser computada numericamente através de uma fórmula de quadratura tipo Gauss da forma: $f_{medio} = \sum_{i=1}^{n} \omega_i f(r_i) + R_n(x)$ determine as abscissas (r_i), os pesos (ω_i) e a expressão do resíduo desta fórmula de quadratura para $n = 2$ e 3.

Problema 6.5 Na tabela abaixo, apresentam-se os valores da viscosidade dinâmica da água a várias temperaturas:

T $(^\circ C)$	μ $\left(\frac{Ns}{m^3}\right) \times 10^3$
0	1,787
5	1,519
10	1,307
20	1,002
30	0,798
40	0,653
50	0,547

O valor médio da viscosidade nesta faixa de temperatura é calculado pela integral: $\mu_{medio} = \frac{1}{50}\int_0^{50} \mu(T)dT$. Aplicando o método de Simpson em cada subintervalo da tabela, determine μ_{medio}.

Problema 6.6 Um foguete é lançado do solo sendo sua aceleração registrada nos 80 primeiros segundos após seu lançamento. Estes valores estão tabelados abaixo:

$t(s)$	0	10	20	30	40	50	60	70	80
$a(m/s^2)$	30,00	31,63	33,44	35,47	37,75	40,33	43,29	46,69	50,67

Baseado nos valores tabelados calcule a velocidade e a altura do foguete ao cabo dos 80 s.

Problema 6.7 Determine os valores das abscissas x_1 e x_2 de modo que a fórmula de quadratura abaixo apresente a maior ordem de acurácia possível:

$$\int_{-1}^{+1} f(x)dx \approx \frac{1}{3}\left[f(-1)+2f(x_1)+3f(x_2)\right]$$

Problema 6.8 Propõe-se a seguinte fórmula de quadratura:

$$\int_0^h f(x)dx = \frac{h}{2}\left[f(0)+f(h)\right]+ah^2\left[f'(0)-f('h)\right]+R_n(x)$$

Calcule o valor da constante a e a expressão do resíduo $R_n(x)$.

Problema 6.9 Determine as abscissas e pesos das fórmulas de quadratura tipo Gauss abaixo:

(a) $\int_0^1 \sqrt{x}f(x)dx \approx \omega_1 f(x_1)+\omega_2 f(x_2)$;

(b) $\int_0^1 \ln(1/x)f(x)dx \approx \omega_1 f(x_1)+\omega_2 f(x_2)$.

Problema 6.10 Na fórmula de quadratura do tipo Lobatto:

$$\int_{-1}^{1} f(x)dx = \omega_1\left[f(-1)+f(+1)\right]+\omega_2\left[f(-\alpha)+f(+\alpha)\right]+\omega_3 f(0)+R_n(x).$$

Calcule o valor da constante α, dos pesos $\omega_1, \omega_2, \omega_3$ e a expressão do resíduo $R_n(x)$.

Problema 6.11 Determine as abscissas, o peso e a expressão do resíduo da fórmula de quadratura tipo Chebyshev:

$$\int_1^0 \frac{1}{\sqrt{x}}f(x)dx = \omega\left[f(x_1)+f(x_2)+f(x_3)\right]+R_n(x).$$

Problema 6.12 Determine as abscissas, os peso e as expressões dos resíduos das fórmulas de quadratura tipo Radau e Lobatto abaixo:

1. $\int_0^1 \frac{1}{\sqrt{x}} f(x)dx = \omega_0 f(0) + \omega_1 f(x_1) + \omega_2 f(x_2) + R_n(x).$
2. $\int_0^1 \frac{1}{\sqrt{x}} f(x)dx = \omega_1 f(x_1) + \omega_2 f(x_2) + \omega_3 f(1) + R_n(x).$
3. $\int_0^1 \frac{1}{\sqrt{x}} f(x)dx = \omega_0 f(0) + \omega_1 f(x_1) + \omega_2 f(x_2) + \omega_3 f(1) + R_n(x).$

Problema 6.13 Determine as abscissas, os pesos e a expressão do resíduo da fórmula de quadratura tipo Gauss para o cômputo de integrais duplas:

$$\int_{-1}^{+1}\int_{-1}^{+1} f(x,y)dydx = \omega_{1,1} f(x_1,y_1) + \omega_{1,2} f(x_1,y_2) + \omega_{2,1} f(x_2,y_1) + \omega_{2,2} f(x_2,y_2) + R_n(x,y)$$

Aplique o método no cálculo da integral: $\int_{-1}^{+1}\int_{-1}^{+1} e^{-(x^2+y^2)}dydx.$

Problema 6.14 Para o cálculo de integrais do tipo: $I = \int_0^\infty e^{-x} f(x)dx$ a quadratura de Gauss-Laguerre[11] é empregada, expressa por:

$I = \sum_{i=0}^{n} \omega_i f(x_i) + \frac{[(n+1)!]^2}{(2n+2)!} f^{(2n+2)}(\xi)$, sendo as abscissas x_i as $(n+1)$ raízes do polinômio da Laguerre de grau $(n+1)$. Na tabela abaixo, listam-se alguns valores das abscissas e dos pesos da quadratura de Gauss-Laguerre.

Raízes (x_i)		Pesos (ω_i)
	Quadratura de Gauss-Laguerre $n=1$	
0,5857864376269		0,8535533905933
3,4142135623731		0,1464466094067
	Quadratura de Gauss-Laguerre $n=2$	
0,415774556659		0,7110930098857
2,294280360462		0,2785177336237
6,289945082879		0,0103892564907
	Quadratura de Gauss-Laguerre $n=3$	
0,3225476886277		0,6031540781881
1,7457609373338		0,357418712579
4,5366205567039		0,0388879155359
9,3950708173346		0,0005392936969

Teste a adequação do procedimento no cômputo da integral: $\int_0^\infty e^{-x}\frac{\text{sen}(x)}{x}dx = \frac{\pi}{4}$, calculando novamente a integral pelos métodos: Simpson, Romberg e Romberg-Lagrange.

Problema 6.15 Calcule numericamente as integrais abaixo:

(a) $\int_0^\infty \frac{e^{-x^2}}{1+x^2}dx;$

(b) $\int_0^\infty \frac{1}{\sqrt{e^x+x}}dx;$

[11] Edmond Nicolas Laguerre (1834-1886).

(c) $\displaystyle\int_0^{\infty} \frac{x}{e^x - 1} dx$;

(d) $\displaystyle\int_0^{\infty} \frac{x}{e^x + 1} dx$;

(e) $\displaystyle\int_0^{\infty} e^{-\left(x+\frac{1}{x}\right)} dx$.

Problema 6.16 A função $Si(x)$ é definida por: $Si(x) = \displaystyle\int_0^x \frac{\text{sen}(\xi)}{\xi} d\xi$. Baseado nesta definição, calcule numericamente a integral: $\displaystyle\int_0^1 \frac{Si(x) - \text{sen}(x)}{x^3} dx$.

Problema 6.17 Aplicando procedimento análogo ao apresentado no Exemplo 6.4 calcule as abscissas e pesos da quadratura nos seguintes casos:

- Quadratura de Gauss para $n = 4$ e 5;
- Quadratura de Gauss-Radau com extremidade inferior para $n = 4$ e 5;
- Quadratura de Gauss-Radau com extremidade superior para $n = 4$ e 5;
- Quadratura de Gauss-Lobatto para $n = 3, 4$ e 5.

7. Resolução Numérica de Equações Diferenciais Ordinárias

7.1 Introdução

Muitos problemas em modelagem de processos químicos são formulados em termos de equações diferenciais. Essas equações diferenciais descrevem a relação entre variáveis dependentes (funções desconhecidas) e suas derivadas em relação às variáveis independentes. Quando há apenas uma variável independente, a equação diferencial correspondente é ordinária e caso haja mais de uma variável independente a equação diferencial é parcial. Os problemas de equações diferenciais ordinárias (EDO) ou de sistemas de equações diferenciais ordinárias classificam-se em dois tipos:

(1) Problema de Valor Inicial (PVI): quando todas as condições de contorno do problema são conhecidas no valor inicial (ou no valor final) da variável independente;

(2) Problema de Valor de Contorno (PVC): quando algumas condições são conhecidas no valor inicial da variável independente e outras no valor final da variável independente. Problemas deste tipo são tratados neste capítulo como problemas de valor inicial em que se buscam as condições iniciais que satisfaçam as condições finais pertinentes.

Os métodos numéricos de discretização empregados para resolver problemas de valor de contorno, tais como diferenças finitas, volumes finitos, elementos finitos e métodos dos resíduos ponderados, estão além do escopo deste texto.

A seguir, apresentam-se exemplos dos dois tipos de problemas descritos por equações diferenciais ordinárias:

■ **Exemplo 7.1** Um tanque de armazenamento de água apresenta uma área de seção transversal de $3{,}0\ m^2$ e em sua saída está instalada uma válvula, conforme Figura 7.1, cuja equação característica é: $q(h) = 8\sqrt{h}$, sendo q a vazão volumétrica em m^3/h e h a altura de água no tanque em metros. Sabe-se que a altura normal de operação do tanque é $4{,}0\ m$. Simule a partida da operação do tanque inicialmente vazio, isto é $h(0) = 0$ e vazão de alimentação do tanque $q_1 = 16 m^3/h$ (constante).

Considerando a massa específica da água como constante, o balanço de massa de água em regime transiente no tanque é descrito, em termos volumétricos, por:

$$\frac{dV(t)}{dt} = q_1(t) - q_2(t),$$

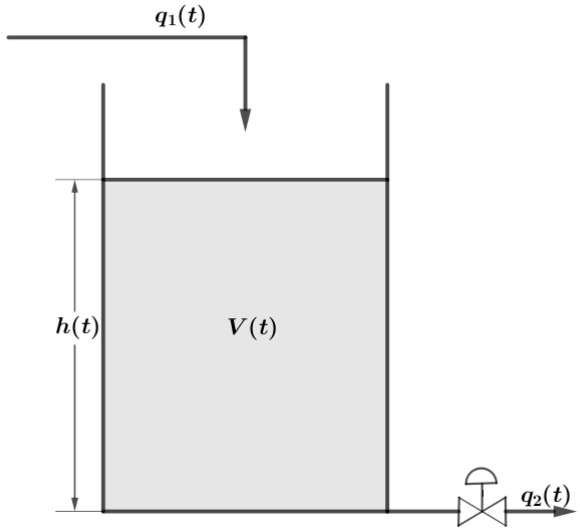


Figura 7.1: Tanque de nível de área constante.

em vista de: $V(t) = A\,h(t)$, $q_1(t) = q_1$ e $q_2(t) = C\sqrt{h}$, obtém-se: $\dfrac{dh(t)}{dt} = \dfrac{q_1 - C\sqrt{h}}{A}$, sujeita à condição inicial $h(0) = 0$.

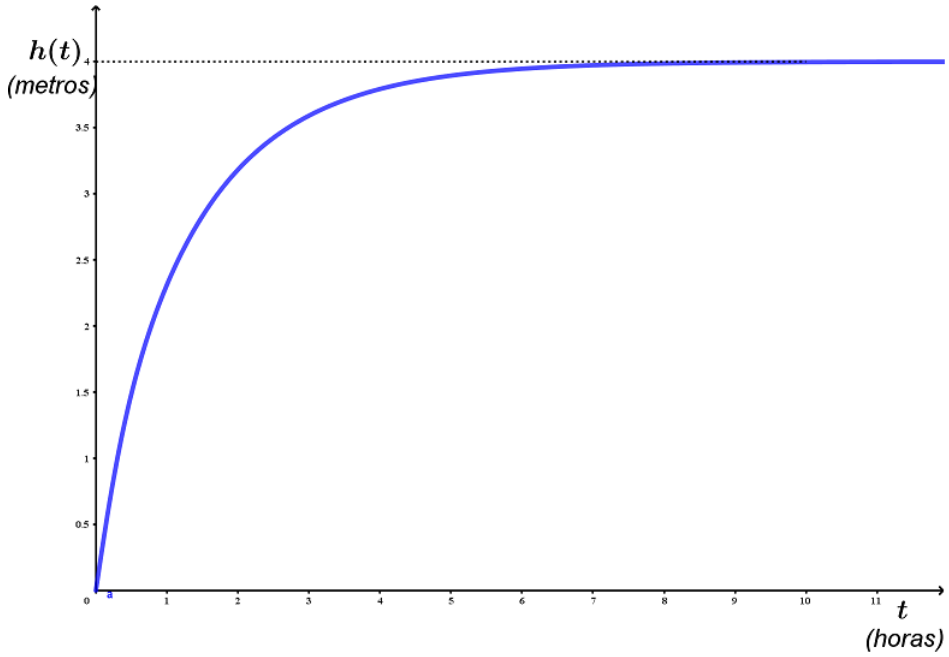


Figura 7.2: Dinâmica do enchimento de um tanque.

Empregando os valores numéricos dos dados e levando em conta que no estado estacionário o nível de líquido no tanque é igual a $\lim\limits_{t \longrightarrow \infty} h(t) = h_{ss} = 4{,}0\,m$ e $q_2 = q_1 \Rightarrow q_1 = 16\,m^3/h = C\sqrt{h_{ss}}$, chega-se finalmente a equação diferencial ordinária não linear: $\dfrac{dh(t)}{dt} = \dfrac{8}{3}\left[2 - \sqrt{h(t)}\right]$, sujeita à condição inicial $h(0) = 0$.

Essa equação diferencial foi resolvida numericamente, por procedimentos descritos no presente capítulo, resultando na forma gráfica representada na Figura 7.2. Verificando-se que o tempo necessário para o nível de água atingir a altura de 4,0 m é de aproximadamente 12 horas.

■

■ **Exemplo 7.2** Deseja-se calcular a distribuição de temperatura em uma aleta fina de seção retangular, conforme mostrado na Figura 7.3.

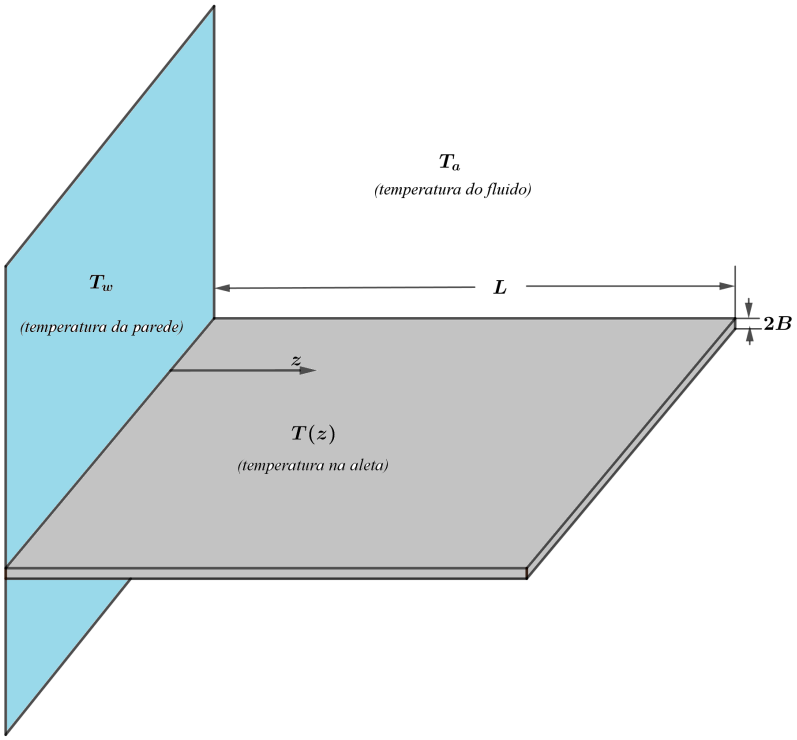


Figura 7.3: Diagrama de uma aleta.

Considerando desprezível a transferência de calor através das áreas laterais da aleta e considerando também a transferência de calor na aleta apenas por condução, tem-se o balanço de energia:

$$\frac{d^2T(z)}{dz^2} = \frac{h}{kB}\left[T(z) - T_a\right] \text{ para } 0 < z < L,$$

sujeita às condições de contorno: $\begin{cases} \text{CC1: condição de contorno na parede } T(0) = T_w; \\ \text{CC2: condição no final da aleta } \left.\dfrac{dT(z)}{dz}\right|_{z=L} = 0, \end{cases}$

sendo B a semi-espessura da aleta, h o coeficiente convectivo de transferência de calor e k a condutividade térmica do material da aleta. A equação diferencial e as condições de contorno são convenientemente reescritas em termos de variáveis e coeficientes adimensionais: $x = \dfrac{z}{L}$, $\theta(x) = \dfrac{T(xL) - T_a}{T_w - T_a}$ e $\alpha = \sqrt{\dfrac{hL^2}{kB}}$, resultando em:

$$\frac{d^2\theta(x)}{dx^2} = \alpha^2\theta(x), \text{ sujeita às condições: } \begin{cases} \text{CC1: } \theta(0) = 1; \\ \text{CC2: } \left.\dfrac{d\theta(x)}{dx}\right|_{x=1} = 0. \end{cases}$$

Definindo as *variáveis de estado* do sistema por: $\begin{cases} y_0(x) = \theta(x) \\ y_1(x) = \dfrac{d\theta(x)}{dx} \end{cases}$,

tem-se, em termos vetoriais:

$$\frac{d\mathbf{y}(x)}{dx} = \begin{pmatrix} \dfrac{dy_0(x)}{dx} \\ \dfrac{dy_1(x)}{dx} \end{pmatrix} = \begin{pmatrix} y_1(x) \\ \alpha^2 y_0(x) \end{pmatrix} \text{ com } y_0(0) = 1 \text{ e } y_1(1) = 0.$$

Deve-se então buscar o valor de $y_1(0)$ que conduza a $y_1(1) = 0$.

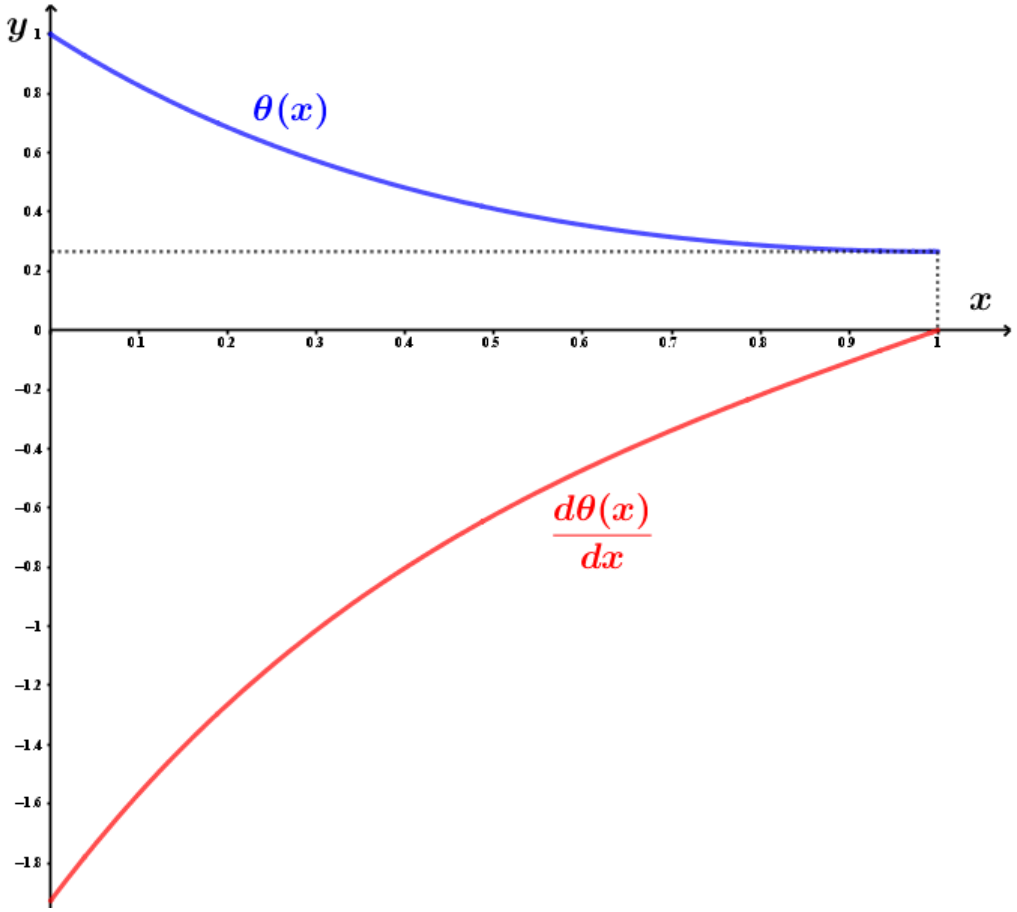


Figura 7.4: Perfis da temperatura e da derivada da temperatura na aleta.

Na realidade este problema, por ser linear e de coeficiente constante, apresenta solução analítica:

$$\begin{cases} \theta(x) = y_0(x) = \dfrac{\cosh[\alpha(1-x)]}{\cosh(\alpha)} \\ \dfrac{d\theta(x)}{dx} = y_1(x) = -\dfrac{\alpha \operatorname{senh}[\alpha(1-x)]}{\cosh(\alpha)} \end{cases}$$

.

Os perfis de $\theta(x)$ e $\dfrac{d\theta(x)}{dx}$, para $\alpha = 2$, encontram-se plotados na Figura 7.4.

Apesar de o problema apresentar solução analítica, o mesmo pode ser empregado para qualificar métodos numéricos empregados em sua resolução. Nesse caso, devido ao fato do problema ser de segunda ordem, homogêneo e linear, aplica-se o *princípio da superposição* estabelecendo que a

solução geral do problema é uma combinação linear de duas soluções linearmente independentes. Assim, seja $\mathbf{y}^{(1)}(x)$ uma solução do problema que satisfaz à condição inicial $\mathbf{y}^{(1)}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ e $\mathbf{y}^{(2)}(x)$ uma solução do problema que satisfaz à condição inicial $\mathbf{y}^{(2)}(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Desse modo,

$$\mathbf{y}(x) = c_1 \mathbf{y}^{(1)}(x) + c_2 \mathbf{y}^{(2)}(x)$$

é uma solução geral do problema original. Os valores de c_1 e c_2 são determinados pelas duas condições de contorno do problema, ou seja: $\begin{cases} y_0(0) = 1 = c_1 \\ y_1(1) = 0 = c_1 y^{(1)}(1) + c_2 y^{(2)}(1) \Rightarrow c_2 = -\dfrac{y^{(1)}(1)}{y^{(2)}(1)} \end{cases}$.

Assim, a solução do problema é: $\mathbf{y}(x) = \mathbf{y}^{(1)}(x) - \dfrac{y^{(1)}(1)}{y^{(2)}(1)} \mathbf{y}^{(2)}(x)$. Esse procedimento foi aplicado, reproduzindo-se o resultado da solução analítica.

Esse mesmo procedimento é aplicável ao problema em que o parâmetro α varia ao longo de x,. O problema continua linear, porém com coeficiente variável, não sendo possível obter uma solução analítica geral. Como ilustração, para caracterizar a natureza compósita do material da aleta considera-se: $\alpha = \begin{cases} 2 \text{ para } 0 < x < 1/2 \\ 1 \text{ para } 1/2 \leq x < 1 \end{cases}$.

Os perfis de $\theta(x)$ e $\dfrac{d\theta(x)}{dx}$ da aleta desse novo problema são apresentados na Figura 7.5.

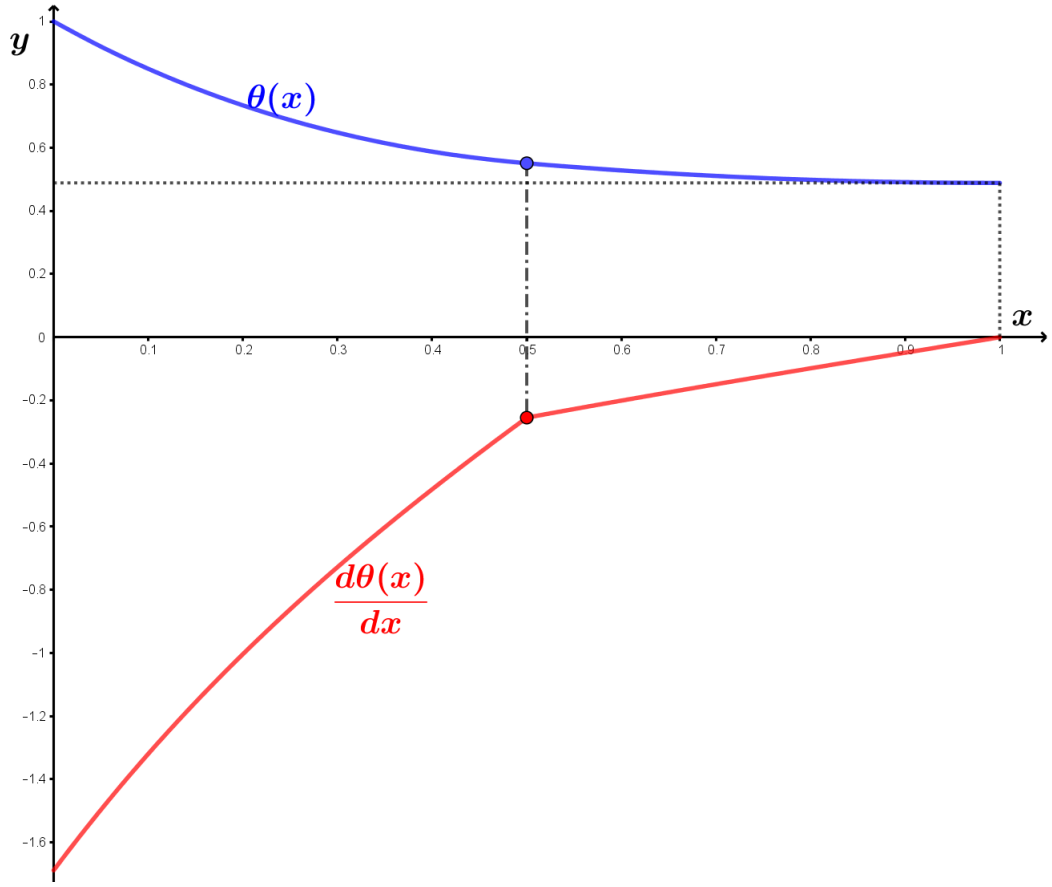


Figura 7.5: Perfis de temperatura e da derivada de temperatura na aleta compósita.

Quando o parâmetro α é função da temperatura, isto é $\alpha = \alpha(\theta)$, o problema torna-se não linear e o princípio da superposição não pode ser mais aplicado. Devendo-se buscar a solução da equação não linear $y_1(1) = f[y_1(0)] = 0$, em que a forma explícita da função f não é conhecida. Para isso, a equação diferencial:

$\frac{d\mathbf{y}(x)}{dx} = \frac{d}{dx}\begin{pmatrix} y_0(x) \\ y_1(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ [\alpha(y_0)]^2 & 0 \end{pmatrix}\begin{pmatrix} y_0(x) \\ y_1(x) \end{pmatrix}$, é resolvida repetidas vezes com a condição inicial: $\begin{pmatrix} y_0(0) \\ y_1(0) \end{pmatrix} = \begin{pmatrix} 1 \\ \xi \end{pmatrix}$, permitindo a construção da curva de $y_1(1)$ *versus* ξ na qual se busca o ponto que intercepta o eixo horizontal, que é o valor de $y_1(0)$, solução do problema.

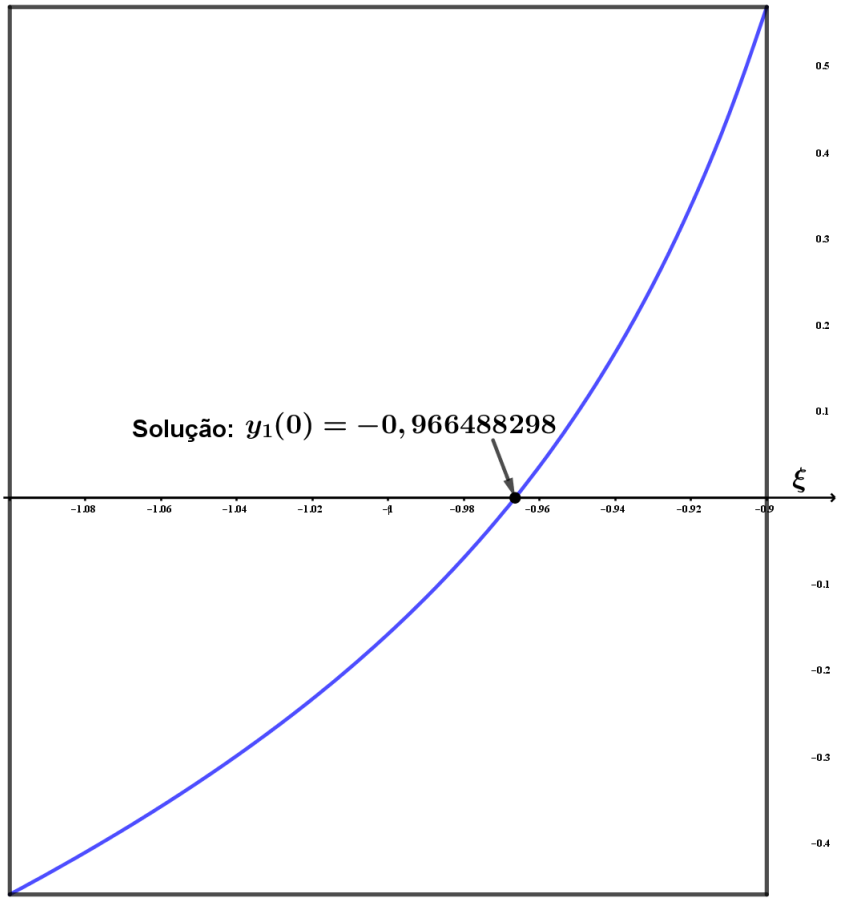


Figura 7.6: Derivada da temperatura na aleta em $x = 1$ para diferentes condições iniciais $y_1(0) = \xi$.

Como ilustração, considera-se $\alpha(\theta) = 2\theta^3 = 2y_0^3$, resultando no problema: $\begin{cases} \frac{dy_0(x)}{dx} = y_1(x) \\ \\ \frac{dy_1(x)}{dx} = 4[y_0(x)]^6 \end{cases}$,

com $\begin{pmatrix} y_0(0) \\ y_1(0) \end{pmatrix} = \begin{pmatrix} 1 \\ \xi \end{pmatrix}$. Analisando a Figura 7.6, em que se plota $y_1(1)$ *versus* ξ, verifica-se que valor de ξ que satisfaz a $y_1(1) = 0$ é igual a -0,966488298, para obter $\mathbf{y}(x)$ resolve-se novamente o problema com a condição inicial: $\begin{pmatrix} y_0(0) \\ y_1(0) \end{pmatrix} = \begin{pmatrix} 1 \\ -0,966488298 \end{pmatrix}$. Os perfis resultantes são qualitativamente semelhantes aos da Figura 7.4. ∎

Através de uma redefinição apropriada das variáveis dependentes do problema, é possível

representar uma equação diferencial ordinária de ordem n através de um sistema de n equações diferenciais ordinárias de primeira ordem. Seja uma EDO de ordem n da forma:

$$\frac{d^n x(t)}{dt^n} = f\left[t, x(t), \frac{dx(t)}{dt}, \frac{d^2x(t)}{dt^2}, \cdots, \frac{d^{n-1}x(t)}{dt^{n-1}}\right], \text{ sujeita às condições iniciais: } \begin{cases} x(0) = x_0 \\ \left.\dfrac{dx(t)}{dt}\right|_{t=0} = x_0' \\ \left.\dfrac{d^2x(t)}{dt^2}\right|_{t=0} = x_0'' \\ \vdots \\ \left.\dfrac{d^{n-1}x(t)}{dt^{n-1}}\right|_{t=0} = x_0^{(n-1)} \end{cases}.$$

Representando esta EDO pelas *n variáveis de estado*: $\begin{cases} x_1(t) = x(t) \\ x_2(t) = \dfrac{dx(t)}{dt} \\ x_3(t) = \dfrac{d^2x(t)}{dt^2} \\ \vdots \\ x_i(t) = \dfrac{d^{i-1}x(t)}{dt^{i-1}} \\ \vdots \\ x_n(t) = \dfrac{d^{n-1}x(t)}{dt^{n-1}} \end{cases}$, resulta:

$$\begin{cases} \dfrac{dx_1(t)}{dt} = \dfrac{dx(t)}{dt} = x_2(t) \\ \dfrac{dx_2(t)}{dt} = \dfrac{d^2x(t)}{dt^2} = x_3(t) \\ \dfrac{dx_3(t)}{dt} = \dfrac{d^3x(t)}{dt^3} = x_4(t) \\ \vdots \\ \dfrac{dx_i(t)}{dt} = \dfrac{d^ix(t)}{dt^i} = x_{i+1}(t) \\ \vdots \\ \dfrac{dx_n(t)}{dt} = \dfrac{d^nx(t)}{dt^n} = f\left[t, x_1(t), x_2(t), \cdots, x_i(t), \cdots, x_n(t)\right] \end{cases},$$

ou, em termos matriciais: $\dfrac{d\mathbf{x}(t)}{dt} = \mathbf{f}[t, \mathbf{x}(t)]$ sujeita às condições iniciais: $, \mathbf{x}(0) = \mathbf{x}_0,$ em que:

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} \text{ e } \mathbf{f}[t, \mathbf{x}(t)] = \begin{pmatrix} x_2(t) \\ x_3(t) \\ \vdots \\ f\left[t, x_1(t), x_2(t), \cdots, x_i(t), \cdots, x_n(t)\right] \end{pmatrix}.$$

Em muitos exemplos de dinâmica de processos químicos essa representação matricial surge naturalmente ao aplicar as leis de conservação de massa, energia e quantidade de movimento, como no exemplo apresentado a seguir.

■ **Exemplo 7.3** Dinâmica de um reator de mistura perfeita não isotérmico.

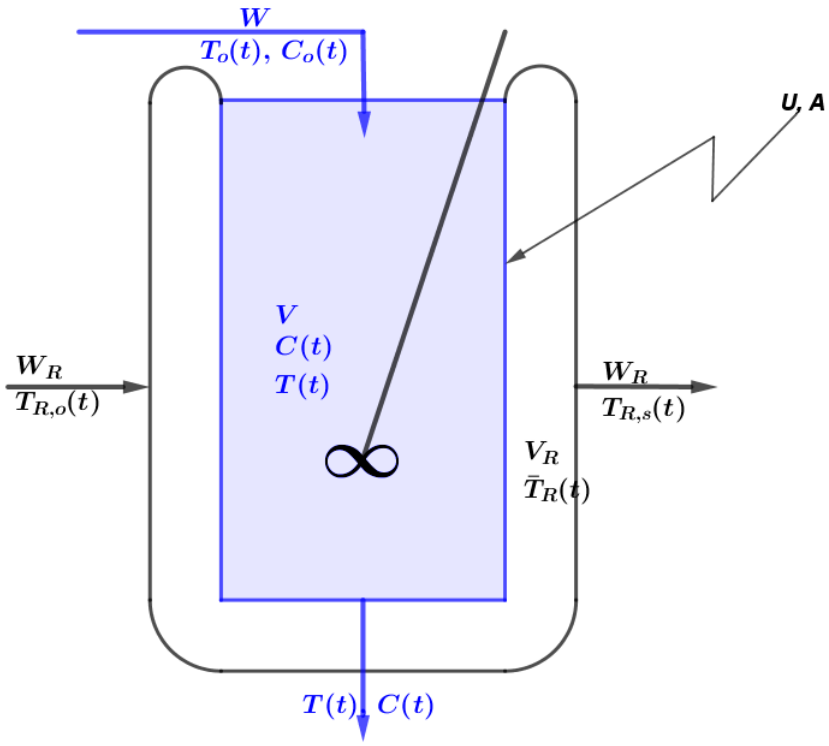


Figura 7.7: Reator de mistura perfeita.

Considerando que ocorre no interior do reator uma reação em fase líquida, exotérmica e irreversível de ordem m, tem-se os balanços:

- Balanço do reagente no interior do reator: $V\dfrac{dC(t)}{dt} = \dfrac{W}{\rho}\left[C_o(t) - C(t)\right] - k_0 e^{-(E/RT)}\left[C(t)\right]^m$
- Balanço de energia no interior do reator:

$$V\rho C_P \frac{dT(t)}{dt} = WC_P\left[T_o(t) - T(t)\right] + k_0 e^{-(E/RT)}\left[C(t)\right]^m\left[-\Delta H_{\text{reação}}\right] - UA\left[T(t) - \bar{T}_R(t)\right]$$

- Balanço de energia no interior da camisa de refrigeração:

$$V_R\rho_r C_{P,R}\frac{d\bar{T}_R(t)}{dt} = W_R C_{P,R}\left[T_{R,o}(t) - T_{R,s}(t)\right] + UA\left[T(t) - \bar{T}_R(t)\right]$$

Considerando, por simplicidade, que o valor médio na temperatura no interior da camisa é a média aritmética entre a sua temperatura de entrada e a de saída, isto é: $\bar{T}_R = \dfrac{T_{R,o} + T_{R,s}}{2} \Rightarrow T_{R,o} - T_{R,s} = 2[T_{R,o} - \bar{T}_R]$ resultando em:

$$V_R\rho_r C_{P,R}\frac{d\bar{T}_R(t)}{dt} = 2W_R C_{P,R}\left[T_{R,o}(t) - \bar{T}_R(t)\right] + UA\left[T(t) - \bar{T}_R(t)\right].$$

Reescrevendo estas equações em termos das variáveis adimensionais:

$$\tau = \frac{t}{(\rho V/W)}; x = \frac{C}{C^*}; y = \frac{T}{T^*}; y_R = \frac{\bar{T}_R}{T^*}; x_o = \frac{C_o}{C^*}; y_o = \frac{T_o}{T^*} \text{ e } y_{R,o} = \frac{T_{R,o}}{T^*},$$

o que dá origem aos parâmetros adimensionais:

$$Da = k_0(C^*)^{m-1}e^{-\gamma}; \gamma = \frac{E}{RT^*}; \beta = \frac{C^*\left(-\Delta H_{\text{reação}}\right)}{\rho V C_P T^*}; \lambda = \frac{UA}{WC_P}; \lambda_R = \frac{UA}{2W_R C_{P,R}}$$

e $\tau_R = \dfrac{[\rho_R V_R/(2W_R)]}{(\rho V/W)}$. Obtêm-se os balanços:

- Balanço do reagente no interior do reator: $\dfrac{dx(\tau)}{d\tau} = x_o(\tau) - x(\tau) - Da \exp\left[-\gamma\left(\dfrac{1}{y(\tau)} - 1\right)\right] x^m(\tau)$
- Balanço de energia no interior do reator:

$$\frac{dy(\tau)}{d\tau} = y_o(\tau) - y(\tau) + \beta Da \exp\left[-\gamma\left(\frac{1}{y(\tau)} - 1\right)\right] x^m(\tau) - \lambda\,[y(\tau) - y_R(\tau)]$$

- Balanço de energia no interior da camisa de refrigeração:

$$\tau_R \frac{dy_R(\tau)}{d\tau} = y_{R,o}(\tau) - y_R(\tau) + \lambda_R\,[y(\tau) - y_R(\tau)]$$

■

Os métodos de resolução numérica de equações diferenciais ordinárias (EDO) que são apresentados neste capítulo são aplicados, sem perda de generalidade, a equações diferenciais ordinárias da forma:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}[t, \mathbf{x}(t)] \text{ para } t > t_0, \text{ sujeita à condição inicial: } \mathbf{x}(t_0) = \mathbf{x}_0,$$

em que $\mathbf{x} \in \mathbb{R}^n$, $t \in \mathbb{R}$ e $\mathbf{f}[t, \mathbf{x}] : \mathbb{R}^{n+1} \to \mathbb{R}^n$.

Teorema 7.1.1 — Teorema de Existência e Unicidade de Solução de uma EDO. Seja $\mathbf{f}(t, \mathbf{x})$ uma função contínua para todo $(t, \mathbf{x})$ em uma região $\mathbb{D} = \{t_0 \le t \le t_{final}, -\infty < ||\mathbf{x}|| < \infty\}$. Além disto, considere continuidade de Lipschitz[1] em $\mathbf{x}$, isto é, existe uma constante real $\mathbb{L} > 0$ tal que para todo $(t, \mathbf{x})$ e $(t, \hat{\mathbf{x}})$ em $\mathbb{D}$,

$$||\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \hat{\mathbf{x}})|| \le \mathbb{L}\,||\mathbf{x} - \hat{\mathbf{x}}||.$$

Então,

(i) Para qualquer $\mathbf{x}_0 \in \mathbb{R}^n$, existe uma solução única $\mathbf{x}(t)$ no intervalo $t_0 \le t \le t_{final}$ para o problema de condição inicial acima. Esta solução é diferenciável.

(ii) A solução $\mathbf{x}(t)$ depende continuamente da condição inicial $\mathbf{x}(t_0) = \mathbf{x}_0$ e se $\hat{\mathbf{x}}(\mathbf{t})$ também for solução da EDO com diferente condição inicial, $\hat{\mathbf{x}}(\mathbf{0}) = \hat{\mathbf{x}}_0$, então

$$||\mathbf{x}(t) - \hat{\mathbf{x}}(t)|| \le e^{\mathbb{L}t}||\mathbf{x}(0) - \hat{\mathbf{x}}(0)||.$$

(iii) Se $\hat{\mathbf{x}}(\mathbf{t})$ satisfaz a EDO submetida a uma perturbação $\mathbf{r}[t, \hat{\mathbf{x}}(t)]$, então

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = \mathbf{f}[t, \hat{\mathbf{x}}(t)] + \mathbf{r}[t, \hat{\mathbf{x}}(t)],$$

em que $\mathbf{r}$ é limitada em $\mathbb{D}$, $||\mathbf{r}|| \le \mathbb{M} > 0$, então

$$||\mathbf{x}(t) - \hat{\mathbf{x}}(t)|| \le e^{\mathbb{L}t}||\mathbf{x}(0) - \hat{\mathbf{x}}(0)|| + \frac{\mathbb{M}}{\mathbb{L}}(e^{\mathbb{L}t} - 1).$$

Nos exemplos apresentados neste capítulo, a **existência** e **unicidade** de solução estão sempre garantidas. No desenvolvimento dos métodos numéricos, visando facilitar a exposição e sem perda de generalidade, o caso escalar é considerado, assim:

$$\frac{dx(t)}{dt} = f[t, x(t)] \text{ para } t > t_0, \text{ sujeita à condição inicial: } x(t_0) = x_0, \tag{7.1}$$

[1] Rudolf Otto Sigismund Lipschitz (1832-1903).

em que: $x \in \mathbb{R}, t \in \mathbb{R}$ e $f[t,x] : \mathbb{R}^2 \to \mathbb{R}$. Para o caso de equações diferenciais ordinárias de ordem superior a um ou n EDOs de primeira ordem, basta considerar a versão vetorial das variáveis e de sua função derivada.

Para o entendimento do desempenho e acurácia dos métodos apresentados neste capítulo, é importante distinguir as três formas da solução do problema:

- **Solução exata**: curva contínua no plano (t,x) que passa por (t_0,x_0) e que satisfaz exatamente a EDO nos demais pontos;
- **Solução numérica**: é um conjunto de pontos discretos no plano (t,x), $[(t_i,u_i)]_{i=0}^{N}$, em que $u_0 = x_0$ e u_i para $i = 1, 2, \cdots, N$ é uma aproximação de $x(t_i)$. A solução numérica é apenas um conjunto discreto de pontos, nada informando sobre valores intermediários;
- **Solução exata no intervalo** i: solução exata do problema no intervalo $t_{i-1} < t \leq t_i$ com a condição inicial $y(t_{i-1}) = u_{i-1}$, isto é, solução de:

$$\frac{dy(t)}{dt} = f[t,y(t)] \text{ para } t_{i-1} < t \leq t_i, \text{ sujeita à condição inicial: } y(t_{i-1}) = u_{i-1}. \quad (7.2)$$

Essas definições permitem classificar os dois tipos de erros na integração de uma EDO:

- **Erro por passo** ou **Erro Local**: é o erro da integração numérica da Equação 7.2 no final do intervalo i, isto é, $\varepsilon_{passo}(t_i) = y(t_i) - u_i$;
- **Erro global**: é o erro da integração numérica da Equação 7.1 no final do intervalo i, isto é, $\varepsilon_{global}(t_i) = x(t_i) - u_i$.

Na Figura 7.8 são ilustradas as soluções exata e numérica e o erro global da integração numérica.

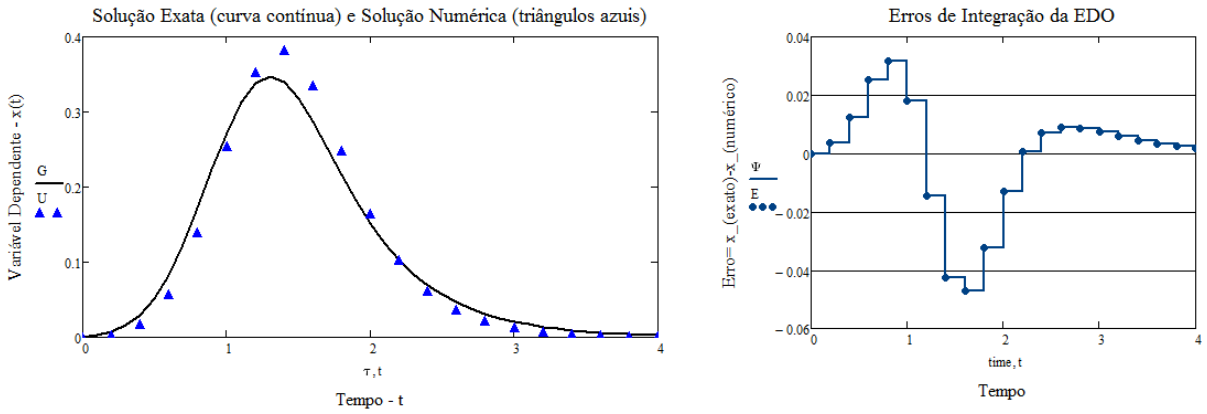


Figura 7.8: Solução exata e solução numérica de uma EDO.

Quando a função f da Equação 7.1 não depende explicitamente de t, diz-se que o sistema é *invariante com o tempo*, o que permite adotar sempre $t_0 = 0$, equivalente a considerar como variável independente o tempo cronometrado a partir de t_0, isto é, a nova variável independente é $(t - t_0)$.

Os métodos numéricos de integração de EDOs podem ser classificados de diferentes formas; classificando-os quanto à dependência a valores anteriores tem-se:

(i) Métodos de Passo Simples: quando o valor da variável dependente no final do intervalo depende apenas de informações do próprio intervalo $[t_{i-1}\, t_i]$, genericamente descrito por: $u_i = g\,[(t_i,u_i),(t_{i-1},u_{i-1})]$;

(ii) Métodos de Passos Múltiplos: quando o valor da variável dependente no final do intervalo depende de informações do intervalo corrente e de intervalos anteriores, representado por: $u_i = g\,[(t_i,u_i),(t_{i-1},u_{i-1}),(t_{i-2},u_{i-2}),\cdots,(t_{i-m},u_{i-m})]$.

Os métodos numéricos de integração de EDOs também podem ser classificados como **explícitos** ou **implícitos** caso o valor da variável dependente **independa** ou **dependa**, respectivamente, de si mesma. Combinando-se essas duas formas de classificação, têm-se os seguintes tipos de métodos:

- Método de passo simples e explícito: $u_i = g\left[t_i,(t_{i-1},u_{i-1})\right]$;
- Método de passo simples e implícito: $u_i = g\left[(t_i,u_i),(t_{i-1},u_{i-1})\right]$;
- Método de passos múltiplos e explícito: $u_i = g\left[t_i,(t_{i-1},u_{i-1}),(t_{i-2},u_{i-2}),\cdots,(t_{i-m},u_{i-m})\right]$;
- Método de passos múltiplos e implícito: $u_i = g\left[(t_i,u_i),(t_{i-1},u_{i-1}),(t_{i-2},u_{i-2}),\cdots,(t_{i-m},u_{i-m})\right]$.

Note que nos métodos implícitos deve se associar ao algoritmo de integração um algoritmo de resolução de equações não lineares (geralmente o método de Newton-Raphson). Desse modo o processo de integração torna-se mais lento, demandando a cada passo de integração o cômputo (analítico ou numérico) da matriz jacobiana do sistema, necessária à aplicação do método de Newton-Raphson. Por outro lado, esses métodos são sempre estáveis, ao contrário dos métodos explícitos que possuem estabilidade condicionada ao tamanho do passo, ou intervalo de integração, como discutido nas próximas seções.

Os métodos podem também ser classificados como de **passo fixo** quando $t_i = t_{i-1} + h$,, sendo h o intervalo de integração, e de **passo variável** quando: $t_i = t_{i-1} + h_i$, isto é, o intervalo de integração h varia com i. Os métodos de passos variável são, via de regra, mais eficientes e robustos, demandando entretanto que ao algoritmo de integração seja acoplado um algoritmo de seleção do tamanho de passo para controle do erro local de integração. Nos métodos descritos a seguir considera-se, por simplicidade, o intervalo de integração como constante, havendo ao final do capítulo uma leve menção aos algoritmos de seleção de passo, assunto este que foge ao escopo do presente texto.

7.2 Métodos de Integração Tipo Euler

Este é o método mais simples e antigo, em sua forma explícita, utilizado na resolução numérica de EDOs que aplicado à resolução da Equação 7.2, é descrito pela equação de diferenças:

$$u_i = u_{i-1} + hf\left[t_{i-1},u_{i-1}\right] \text{ para } i = 1,2\cdots,N, \text{ em que } u_0 = x_0, t_{i-1} = t_i + h \text{ e } h = \frac{t_{final}}{N}.$$

O **método de Euler explícito** pode ser interpretado de três formas distintas:

(a) Diferenças Finitas: aproximando no intervalo i, em que $t_{i-1} \leq t < t_i$, a derivada contínua por: $\dfrac{dy(t)}{dt} \approx \dfrac{u_i - u_{i-1}}{h}$ e igualando esta aproximação ao valor da derivada no início do intervalo (*método explícito*), chega-se a:
$\dfrac{u_i - u_{i-1}}{h} = f(t_{i-1},u_{i-1}) \Rightarrow u_i = u_{i-1} + hf\left[t_{i-1},u_{i-1}\right]$.

(b) Aproximação Linear de $y(t)$: aproximando $y(t)$ pela reta (Figura 7.9):
$y(t) \approx y(t)|_{t_{i-1}} + \left.\dfrac{dy(t)}{dt}\right|_{t_{i-1}} (t - t_{i-1}) = u_{i-1} + f(t_{i-1},u_{i-1})(t - t_{i-1})$, assim:
$y(t_i) \approx u_i = u_{i-1} + f(t_{i-1},u_{i-1})(t_i - t_{i-1}) = u_{i-1} + hf(t_{i-1},u_{i-1})$.

(c) Por Integração Retangular: a integração de ambos os lados da Equação 7.2 de t_{i-1} a t_i conduz a:
$\displaystyle\int_{t_{i-1}}^{t_i} \frac{dy(t)}{dt}dt = y(t_i) - y(t_{i-1}) = \int_{t_{i-1}}^{t_i} f[t,y(t)]dt$, mas: $y(t_{i-1}) = u_{i-1}$ e considerando $f[t,y(t)] \approx f[t_{i-1},u_{i-1}]$ em todo o intervalo $[t_{i-1},t_i)$ (área do retângulo na Figure 7.10), resulta: $y(t_i) \approx u_i = u_{i-1} + f(t_{i-1},u_{i-1})(t_i - t_{i-1}) = u_{i-1} + hf(t_{i-1},u_{i-1})$.

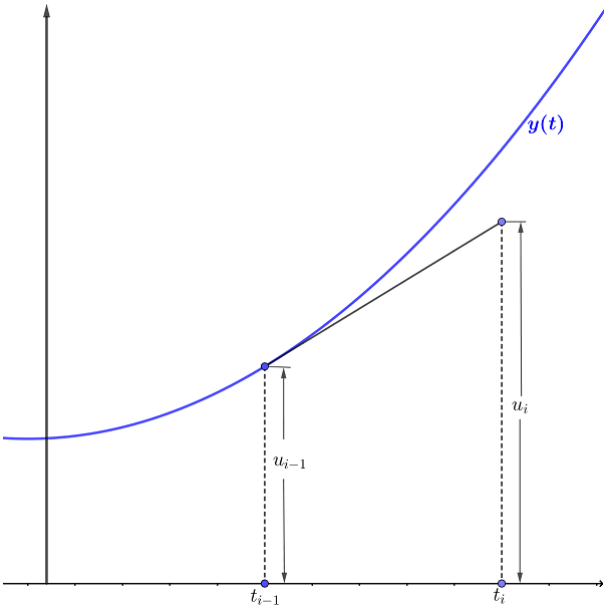


Figura 7.9: Método de Euler explícito por linearização.

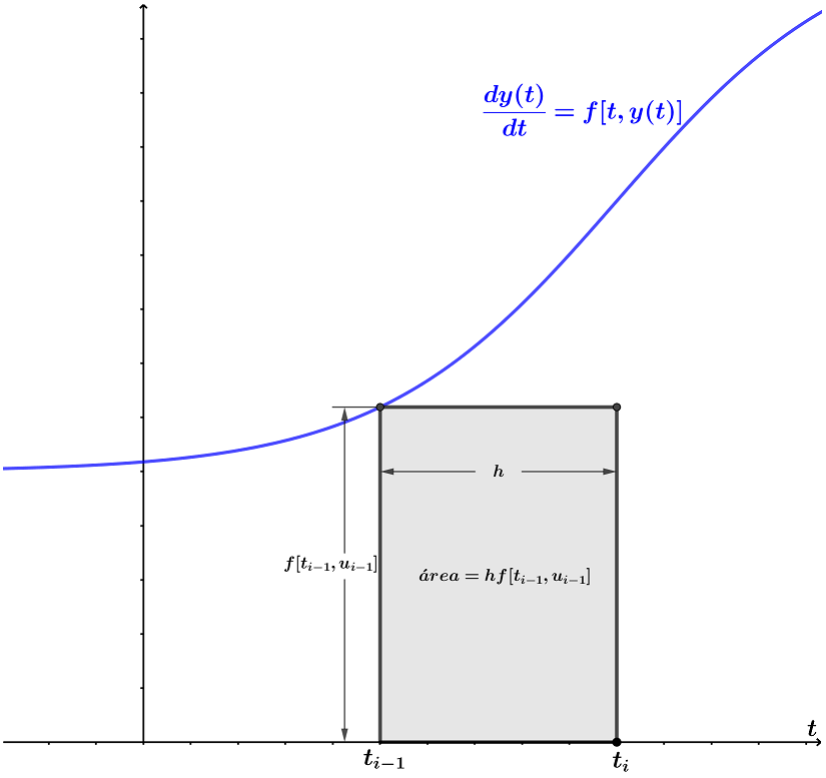


Figura 7.10: Método de Euler explícito por integração retangular.

Modificando-se o método de Euler, interpretado no intervalo $t_{i-1} \leq t < t_i$, como um método de diferenças finitas em que a derivada contínua é aproximada por: $\frac{dy(t)}{dt} \approx \frac{u_i - u_{i-1}}{h}$ e igualando

esta aproximação ao valor da derivada no **final** do intervalo (*método implícito*), chega-se a:

$$u_i = u_{i-1} + hf[t_i, u_i] \text{ para } i = 1, 2 \cdots, N, \text{ em que } u_0 = x_0, t_{i-1} = t_i + h \text{ e } h = \frac{t_{final}}{N}.$$

Este procedimento é o **método de Euler implícito** que demanda, em cada intervalo de integração, a utilização de um algoritmo de resolução de equação não linear se a função $f(t,x)$ for não linear.

■ **Exemplo 7.4** Aplicando o método de Euler a EDO de primeira ordem, linear, de coeficiente constante e homogênea: $\frac{dx(t)}{dt} = -\alpha x(t)$, em que $\alpha > 0$ e $x(0) = 1$. A solução analítica desta equação é $x(t) = e^{-\alpha t}$ e $f[t, x(t)] = -\alpha x(t)$.

A equação de diferenças resultante da aplicação do método de Euler explícito à equação é $u_i = u_{i-1} - h\alpha u_{i-1} = (1 - \alpha h)u_{i-1}$ com $u_0 = 1$.

Definindo $q = (1 - \alpha h) =$ constante, identifica-se esta equação de diferenças como uma **progressão geométrica** de razão q e primeiro termo igual à 1, cuja solução é $u_i = q^i$ para $i = 0, 1, 2, \cdots, N$.

Desse modo, este procedimento só será convergente se $|q| < 1$, como $\alpha h > 0 \Rightarrow q = 1 - \alpha h < 1$,

havendo três possibilidades: $\begin{cases} h > \frac{2}{\alpha} \Rightarrow q < -1 \text{ processo não convergente e oscilatório} \\ \frac{1}{\alpha} < h < \frac{2}{\alpha} \Rightarrow -1 < q < 0 \text{ processo convergente e oscilatório} \\ h < \frac{1}{\alpha} \Rightarrow 0 < q < 1 \text{ processo convergente e não-oscilatório} \end{cases}$.

A equação de diferenças resultante da aplicação do método de Euler implícito à equação é $u_i = u_{i-1} - h\alpha u_i \Rightarrow u_i = \left(\frac{1}{1+\alpha h}\right) u_{i-1}$ com $u_0 = 1$.

Definindo $q = \frac{1}{1+\alpha h}$ =constante, identifica-se esta equação de diferenças como uma **progressão geométrica** de razão q e primeiro termo igual à unidade, cuja solução é $u_i = q^i$ para $i = 0, 1, 2, \cdots, N$. Desse modo, este procedimento sempre será convergente e não-oscilatório pois $0 < q < 1$.

Na Figura 7.11 são ilustrados todos esses comportamentos das soluções aproximadas, comparados à solução exata.

■

A caracterização da acurácia dos métodos de resolução numérica de EDOs é feita através da expansão em série de Taylor da função $y(t)$ em torno do ponto t_{i-1}, assim:

$$y(t) = y(t_{i-1}) + \left.\frac{dy(t)}{dt}\right|_{t_{i-1}} (t - t_{i-1}) + \frac{1}{2!}\left.\frac{d^2y(t)}{dt^2}\right|_{t_{i-1}} (t - t_{i-1})^2 + \cdots + \frac{1}{n!}\left.\frac{d^ny(t)}{dt^n}\right|_{t_{i-1}} (t - t_{i-1})^n + R_n(t),$$

em que $R_n(t) = \frac{1}{(n+1)!}\left.\frac{d^{n+1}y(t)}{dt^{n+1}}\right|_{t=\xi} (t - t_{i-1})^{(n+1)}$.

Adotando $t = t_i \Rightarrow t - t_{i-1} = h$ e em vista de $y(t_{i-1}) = u_{i-1}$, obtém-se

$$y(t_i) = u_{i-1} + \left.\frac{dy(t)}{dt}\right|_{t_{i-1}} h + \frac{1}{2!}\left.\frac{d^2y(t)}{dt^2}\right|_{t_{i-1}} h^2 + \cdots + \frac{1}{n!}\left.\frac{d^ny(t)}{dt^n}\right|_{t_{i-1}} h^n + R_n(t),$$

em que $R_n(t) = \frac{1}{(n+1)!}\left.\frac{d^{n+1}y(t)}{dt^{n+1}}\right|_{t=\xi} h^{n+1}$.

Partindo da Equação 7.2, $\frac{dy(t)}{dt} = f[t, y(t)] \Rightarrow \left.\frac{dy(t)}{dt}\right|_{t_{i-1}} = f[t_{i-1}, u_{i-1}] = f_{i-1}$.

$\frac{d^2y(t)}{dt^2} = \frac{\partial f(t,y)}{\partial t} + \frac{\partial f(t,y)}{\partial y}\frac{dy}{dt} = \frac{\partial f}{\partial t} + f\frac{\partial f}{\partial y} = f_t + ff_y$, logo, $\left.\frac{d^2y(t)}{dt^2}\right|_{t_{i-1}} = f_{t,i-1} + f_{i-1}f_{y,i-1}$.

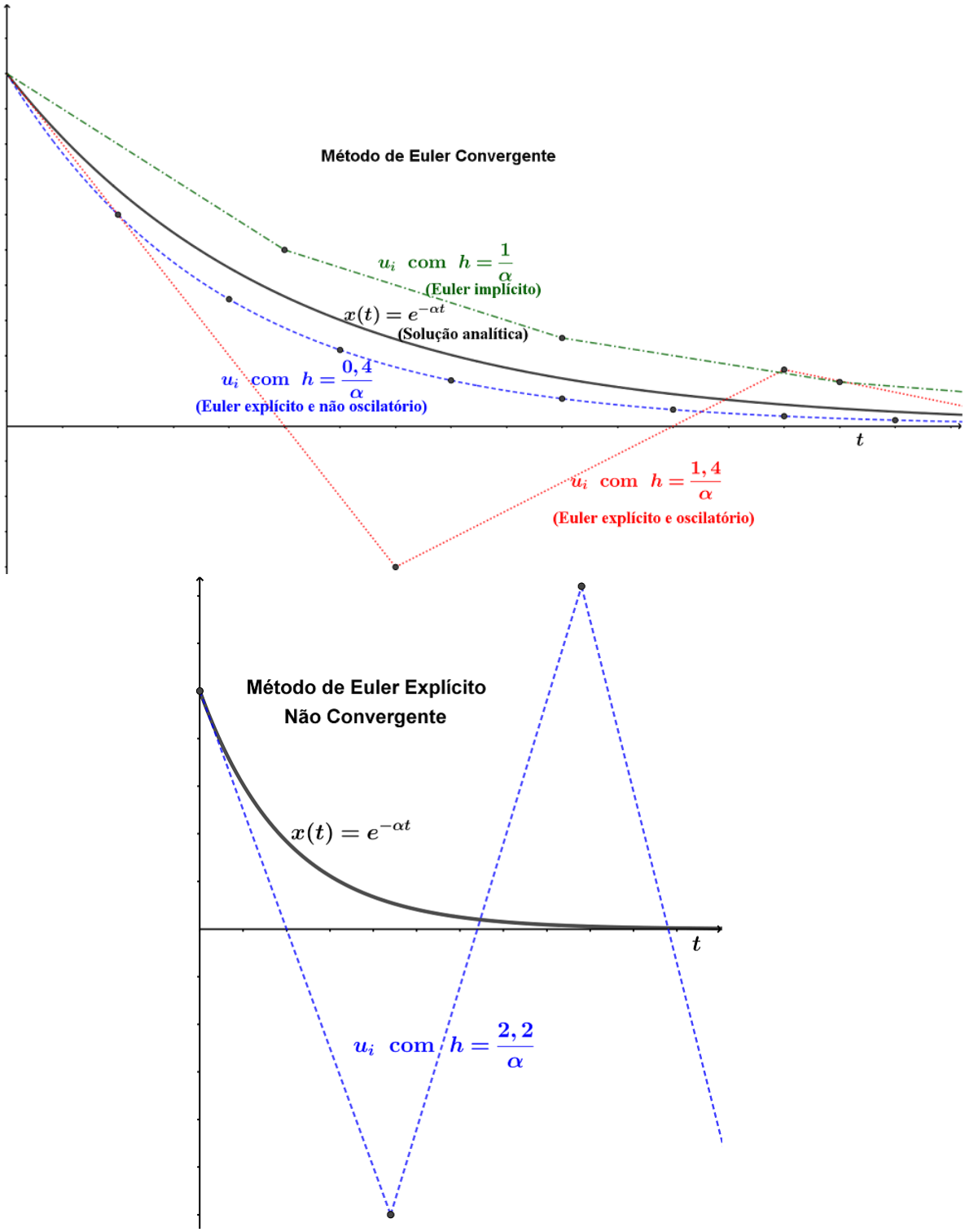


Figura 7.11: Método de Euler para EDO de primeira ordem linear e homogênea.

Permitindo identificar o operador $\frac{d}{dt} \equiv \frac{\partial}{\partial t} + f\frac{\partial}{\partial y}$.

$$\frac{d^3y(t)}{dt^3} = \left(\frac{\partial}{\partial t} + f\frac{\partial}{\partial y}\right)\left(\frac{\partial f}{\partial t} + f\frac{\partial f}{\partial y}\right) = f_{tt} + 2ff_{ty} + f^2 f_{yy} + (f_t + ff_y)f_y,$$

logo $\left.\dfrac{d^3y(t)}{dt^3}\right|_{t_{i-1}} = f_{tt,i-1} + 2f_{i-1}f_{ty,i-1} + f^2 f_{yy,i-1} + (f_{t,i-1} + f_{i-1}f_{y,i-1})f_{y,i-1}$.

Resultando em:

$$y(t_i) = u_{i-1} + f_{i-1}h + \frac{1}{2!}(f_{t,i-1} + f_{i-1}f_{y,i-1})h^2 + \\ + \frac{1}{3!}[f_{tt,i-1} + 2f_{i-1}f_{ty,i-1} + f^2 f_{yy,i-1} + (f_{t,i-1} + f_{i-1}f_{y,i-1})f_{y,i-1}]h^3 + \vartheta[h^4].$$

Em que $\vartheta[h^4]$ designa termos de ordem igual e superior a h^4.

A acurácia dos métodos de resolução numérica de EDOs é verificada confrontando a expansão em série de Taylor do método em torno de t_{i-1} com a expansão acima. Esta verificação é exemplificada a seguir nos dois métodos de Euler (explícito e implícito).

- Método de Euler Explícito $u_i = u_{i-1} + hf[t_{i-1}, u_{i-1}]$ que reproduz exatamente os dois primeiros termos da expansão em Série de Taylor de $y(t)$, permitindo concluir que o erro/passo ou erro local do método é da ordem de h^2, isto é $\varepsilon_{passo}(t_i) = y(t_i) - u_i = \mathbf{C}h^2$;
- Método de Euler Implícito $u_i = u_{i-1} + hf[t_i, u_i]$, expandindo a função $f[t_i, u_i]$ em torno de (t_{i-1}, u_{i-1}), obtém-se
  $f[t_i, u_i] = f|_{t_{i-1},u_{i-1}} + \left.\dfrac{\partial f}{\partial t}\right|_{t_{i-1},u_{i-1}} h + \left.\dfrac{\partial f}{\partial u}\right|_{t_{i-1},u_{i-1}} (u_i - u_{i-1}) = f_{i-1} + f_{t,i-1}h + f_{y,i-1}(u_i - u_{i-1})$.
  $u_i = u_{i-1} + f_{i-1}h + f_{t,i-1}h^2 + f_{y,i-1}(u_i - u_{i-1})h$, considerando a expansão de
  $u_i = u_{i-1} + a_1 h + a_2 h^2 + \cdots \Rightarrow u_i - u_{i-1} = a_1 h + a_2 h^2 + \cdots$, assim
  $u_i = u_{i-1} + f_{i-1}h + [f_{t,i-1} + a_1 f_{y,i-1}]h^2 + \cdots = u_{i-1} + a_1 h + a_2 h^2 + \cdots$, comparando os termos de mesma potência de h resulta $a_1 = f_{i-1}$ e $a_2 = f_{t,i-1} + f_{i-1}f_{y,i-1}$.
  Então $u_i = u_{i-1} + f_{i-1}h + (f_{t,i-1} + f_{i-1}f_{y,i-1})h^2 + \cdots$ reproduzindo os dois primeiros termos da expansão em Série de Taylor de $y(t)$, permitindo concluir que o erro/passo do método é também da ordem de h^2, isto é $\varepsilon_{passo}(t_i) = y(t_i) - u_i = \mathbf{C}h^2$.

Verificando-se que os dois métodos de Euler apresentados (implícito e explícito) apresentam o erro/passo de segunda ordem.

Deve-se enfatizar que apenas no primeiro passo de integração a condição inicial de $y(t)$ na Equação 7.2 é igual à condição inicial da solução exata, $x(t)$ da Equação 7.1, o que implica em $y(t_1) = x(t_1)$, nos passos seguintes tal igualdade não mais existe pois $y(t)$ utiliza uma condição inicial *inexata* u_{i-1}.

Antes de particularizar para os métodos de Euler, considera-se um método numérico de resolução de EDOs que apresenta $|\varepsilon_{passo}(t_i)| = \mathbf{C}_i h^{m+1}$, isto é, apresenta um erro de ordem $(m+1)$ por passo, assim:

Passo	Erro/passo
Primeiro	$\lvert\varepsilon_{passo}(t_1)\rvert = \lvert y(t_1) - u_1\rvert = \mathbf{C}_1 h^{m+1}$
Segundo	$\lvert\varepsilon_{passo}(t_2)\rvert = \lvert y(t_2) - u_2\rvert = \mathbf{C}_2 h^{m+1}$
⋮⋮⋮	⋮⋮⋮
i-ésimo	$\lvert\varepsilon_{passo}(t_i)\rvert = \lvert y(t_1) - u_i\rvert = \mathbf{C}_i h^{m+1}$
⋮⋮⋮	⋮⋮⋮
N-ésimo	$\lvert\varepsilon_{passo}(t_N)\rvert = \lvert y(t_N) - u_N\rvert = \mathbf{C}_N h^{m+1}$
Soma $\left(\lvert\varepsilon_{global}\rvert\right)$	$\left(\sum_{i=1}^{N} \mathbf{C}_i\right) h^m \leq N\,\mathbf{C}_{max} h^{m+1} = \mathbf{C}_{max} t_{final} h^m = \mathbf{C}^* h^m$

Assim, *via de regra*, se o método é de ordem $(m+1)$ por passo, o erro acumulado após N passos é de ordem m. Como os métodos de Euler explícito e implícito são de segunda ordem por passo, globalmente são **métodos de primeira ordem**.

Uma forma bastante empregada para aumentar a acurácia dos métodos é o emprego da *Extrapolação de Richardson*, descrita na Seção 6.2.1. Quando aplicada ao método de Euler explícito dá origem a uma forma aprimorada do procedimento com um erro por passo da ordem de h^3, conforme descrito a seguir.

- Integrando a EDO com um passo h resulta $u_i^{(1)} = u_{i-1} + hf(t_{i-1,u_{i-1}})$
- Integrando EDO com dois passos $h/2$ resulta $\begin{cases} u_{i-1/2} = u_{i-1} + \dfrac{h}{2} f(t_{i-1,u_{i-1}}) \\ u_i^{(2)} = u_{i-1/2} + \dfrac{h}{2} f(t_{i-1/2}, u_{i-1/2}) \end{cases}$

Aplicando a extrapolação de Richardson baseada no fato do erro global do método de Euler ser de primeira ordem, obtém-se
$u_i = 2u_i^{(2)} - u_i^{(1)} = 2\left[u_{i-1/2} + \dfrac{h}{2} f(t_{i-1/2}, u_{i-1/2})\right] - \left[u_{i-1} + hf(t_{i-1,u_{i-1}})\right]$, e tendo em vista que $2u_{i-1/2} = 2u_{i-1} + hf(t_{i-1,u_{i-1}})$, resulta $u_i = u_{i-1} + hf(t_{i-1/2}, u_{i-1/2})$.
Dando origem ao **método de Euler modificado** que apresenta um erro global de segunda ordem e **dois estágios** por passo como descrito a seguir:

$$\begin{cases} \text{Primeiro estágio: } u_{i-1/2} = u_{i-1} + \dfrac{h}{2} f(t_{i-1}, u_{i-1}) \\ \text{Segundo estágio: } u_i = u_{i-1} + hf(t_{i-1/2}, u_{i-1/2}) \end{cases}$$

Esta modificação pode ser expressa na forma genérica:

$$\begin{cases} \text{Primeiro estágio: } g_1 = hf(t_{i-1}, u_{i-1}) \\ \text{Segundo estágio: } g_2 = hf(t_{i-1} + ch, u_{i-1} + cg_1) \end{cases} \Rightarrow u_i = u_{i-1} + (\omega_1 g_1 + \omega_2 g_2)$$

No método de Euler modificado tem-se $\omega_1 = 0$, $\omega_2 = 1$ e $c = 1/2$. No caso geral, os coeficientes ω_1, ω_2 e c são determinados de modo a garantir que o erro/passo seja de terceira ordem, isto é deve equivaler à expansão:

$$y(t_i) = u_{i-1} + f_{i-1}h + \frac{1}{2!}(f_{t,i-1} + f_{i-1}f_{y,i-1})h^2 + \vartheta[h^3].$$

Expandindo g_2 em série de Taylor em torno de (t_{i-1}, u_{i-1}):
$g_2 = f_{i-1}h + ch^2\left[f_{t,i-1} + f_{i-1}f_{y,i-1}\right] + \vartheta[h^2]$ e em vista de $g_1 = hf_{i-1}$ resulta
$u_i = u_{i-1} + (\omega_1 + \omega_2) f_{i-1}h + \omega_2(f_{t,i-1} + f_{i-1}f_{y,i-1})ch^2$, comparando esta expansão com a expansão de $y(t_i)$ obtém-se: $\begin{cases} \omega_1 + \omega_2 = 1 \\ \omega_2 c = 1/2 \end{cases}$, estas são as chamadas **equações de ordem** do método e qualquer valor de ω_1, ω_2 e c que as satisfaçam garantem que o erro/passo é de ordem de h^3 o que resulta em um erro acumulado após $n[>1]$ passos de integração de ordem de h^2.

Uma possível solução seria $\begin{cases} \omega_1 = 0 \text{ e } \omega_2 = 1 \\ c = 1/2 \end{cases}$, outra solução possível seria: $\begin{cases} \omega_1 = \omega_2 = 1/2 \\ c = 1 \end{cases}$.

Dando origem aos métodos

(1) Método de Euler modificado

$$\begin{cases} \text{Primeiro estágio: } g_1 = hf(t_{i-1}, u_{i-1}) \\ \text{Segundo estágio: } g_2 = hf\left(t_{i-1} + \dfrac{h}{2}, u_{i-1} + \dfrac{g_1}{2}\right) \end{cases} \Rightarrow u_i = u_{i-1} + g_2.$$

(2) Método de Euler aprimorado

$$\begin{cases} \text{Primeiro estágio: } g_1 = hf(t_{i-1}, u_{i-1}) \\ \text{Segundo estágio: } g_2 = hf(t_{i-1}+h, u_{i-1}+g_1) \end{cases} \Rightarrow u_i = u_{i-1} + \frac{g_1+g_2}{2}.$$

Verifica-se que o método de Euler implícito, com um estágio, descrito por:

$$g_1 = hf\left(t_{i-1}+\frac{h}{2}, u_{i-1}+\frac{g_1}{2}\right) \Rightarrow u_i = u_{i-1}+g_1,$$

é um método de segunda ordem e sempre estável, pois:

$g_1 = hf\left(t_{i-1}+\frac{h}{2}, u_{i-1}+\frac{g_1}{2}\right) = h\left[f_{i-1}+\frac{1}{2}(hf_{t,i-1}+g_1 f_{y,i-1})\right]+\vartheta[h^3]$, em vista de

$g_1 = hf_{i-1}+a_2h^2+\cdots \Rightarrow g_1 = h\left[f_{i-1}+\frac{h}{2}(f_{t,i-1}+f_{i-1}f_{y,i-1})\right]+\vartheta[h^3]$.

Então $u_i = u_{i-1}+f_{i-1}h+\frac{1}{2!}(f_{t,i-1}+f_{i-1}f_{y,i-1})h^2+\vartheta[h^3]$. Comprovando-se que é um método de terceira ordem por passo e de segunda ordem globalmente.

Este último método pode ser deduzido a partir da aproximação linear de $y(t)$ no intervalo i de acordo com $y(t) \approx y^{(1)}(t) = u_{i-1} + \left(\frac{t-t_{i-1}}{h}\right) g_1$. A seguir, define-se o *Resíduo* desta aproximação por $R^{(1)}(t) = \frac{dy^{(1)}(t)}{dt} - f[t, y^{(1)}(t)] = \frac{g_1}{h} - f\left[t, u_{i-1}+\left(\frac{t-t_{i-1}}{h}\right)g_1\right]$.

Calculando-se g_1 tal que $\int_{t_{i-1}}^{t_i} R^{(1)}(t) = 0$, isto é $g_1 = \int_{t_{i-1}}^{t_i} f\left[t, u_{i-1}+\left(\frac{t-t_{i-1}}{h}\right)g_1\right] dt$.

Esta última integral é calculada por quadratura de Gauss com um ponto, que é o ponto central do intervalo, resultando em: $g_1 = hf\left(t_{i-1}+\frac{h}{2}, u_{i-1}+\frac{g_1}{2}\right)$, expressão análoga à anteriormente obtida. Verificando-se que esta expressão equivale a anular o resíduo da EDO da Equação 7.2 no ponto central do intervalo.

■ **Exemplo 7.5** Aplicando esses últimos métodos à EDO dos exemplos anteriores: $\frac{dx(t)}{dt} = -\alpha x(t)$, em que $\alpha > 0$ e $x(0) = 1$, e cuja solução analítica é $x(t) = e^{-\alpha t}$ e $f[t, x(t)] = -\alpha x(t)$, tem-se:

(1) Método de Euler modificado

$$\begin{cases} \text{Primeiro estágio: } g_1 = -h\alpha u_{i-1} \\ \text{Segundo estágio: } g_2 = -h\alpha\left(1-\frac{h\alpha}{2}\right)u_{i-1} \end{cases} \Rightarrow u_i = \left(1-\alpha h+\frac{\alpha^2h^2}{2}\right)u_{i-1}.$$

(2) Método de Euler aprimorado

$$\begin{cases} \text{Primeiro estágio: } g_1 = -h\alpha u_{i-1} \\ \text{Segundo estágio: } g_2 = -h\alpha(1-h\alpha)u_{i-1} \end{cases} \Rightarrow u_i = \left(1-\alpha h+\frac{\alpha^2h^2}{2}\right)u_{i-1}.$$

Método de Euler implícito, com um estágio, descrito por:

$$g_1 = -h\alpha\left(u_{i-1}+\frac{g_1}{2}\right) \Rightarrow g_1 = \frac{\alpha h}{1+\alpha h/2}u_{i-1}, \text{ então } u_i = \left(\frac{1-\alpha h/2}{1+\alpha h/2}\right)u_{i-1}.$$

Definindo-se como o fator multiplicador do método a razão $\mu(h) = \dfrac{u_i}{u_{i-1}}$, sendo seu valor exato, neste exemplo, $\mu_{exato}(h) = e^{-\alpha h}$. Deste modo, a lei de formação destes métodos é de uma progressão geométrica de razão igual a $\mu(h)$, que para ser convergente deve apresentar $|\mu(h)| < 1$. Os dois métodos de Euler explícitos com dois estágios apresentam o mesmo fator multiplicador e igual a $\mu(h) = \left(1 - \alpha h + \dfrac{\alpha^2 h^2}{2}\right)$ para estes métodos serem convergentes é necessário que $h < \dfrac{2}{\alpha}$ além disto, neste intervalo, $\dfrac{1}{2} \leq \mu(h) < 1$ (estável e não-oscilatório).

O método de Euler implícito, com um estágio, apresenta $\mu(h) = \left(\dfrac{1-\alpha h/2}{1+\alpha h/2}\right)$ que apresenta $0 < \mu(h) < 1$ se $h < \dfrac{2}{\alpha}$ (estável e não-oscilatório) e $-1 < \mu(h) < 0$ se $h > \dfrac{2}{\alpha}$ (estável e oscilatório). ■

Outra modificação do método de Euler é o **método de Crank[2]-Nicolson[3]** (ou trapézios), que é um método implícito com erro local de ordem de h^3 , portanto um método implícito de segunda ordem, e tem a forma:

$$u_i = u_{i-1} + \frac{h}{2}[f(t_i, u_i) + f(t_{i-1}, u_{i-1})].$$

Que pode ser prontamente verificada pela expansão de $y(t)$ e $f[t, y(t)]$ em séries de Taylor em torno de $t_{1-1/2} = t_i - h/2$, e então subtraindo $y(t_{i-1})$ de $y(t_i)$ e adicionado $f[t_{i-1}, y(t_{i-1})]$ a $f[t_i, y(t_i)]$ para eliminar $f[t_{i-1/2}, y(t_{i-1/2})]$ da expansão de $y(t)$.

7.3 Métodos de Integração Tipo Runge-Kutta

Os métodos de integração tipo Runge-Kutta[4] são aperfeiçoamentos dos métodos de Euler, obtendo-se erros de maiores ordem sem a necessidade do cômputo de derivadas superiores de $y(t)$. Na realidade o método de Euler simples pode também ser classificado como um método de Runge-Kutta de primeira ordem e as modificações apresentadas na Seção 7.2: método de Euler modificado, método de Euler aprimorado, método de Euler implícito com 1 estágio e método de Crank-Nicolson são métodos de Runge-Kutta de segunda ordem. De um modo geral, os métodos de Runge-Kutta com ν estágios podem ser expressos por:

$$g_i = h f\left[t_{i-1} + c_i h,\, u_{i-1} + \sum_{j=1}^{\nu} a_{i,j} g_j\right]$$

$$u_i = u_{i-1} + \sum_{j=1}^{\nu} \omega_j g_j$$

Os coeficientes $c_i, a_{i,j}$ e ω_j são apresentados na forma tabular em um arranjo proposto por Butcher (1996).

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,\nu} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,\nu} \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ c_\nu & a_{\nu,1} & a_{\nu,2} & \cdots & a_{\nu,\nu} \\ \hline & \omega_1 & \omega_2 & \cdots & \omega_\nu \end{array}$$

[2]John Crank (1916-2006).
[3]Phyllis Nicolson (1917-1968).
[4]Martin Wilhelm Kutta (1867-1944).

Se $a_{i,j} = 0$ para $i \leq j$ (na diagonal e acima da diagonal da matriz **A** só há termos nulos) o método é **explícito**, caso $a_{i,j} = 0$ para $i < j$ (acima da diagonal da matriz **A** só há termos nulos) o método é dito **semi-implícito** e, caso exista algum termo não nulo acima da diagonal de **A**, o método é **implícito**.

A determinação dos coeficientes c_i, $a_{i,j}$ e ω_i é feita após estabelecidos o número de estágios ν e a ordem da acurácia desejada para deduzir então as **equações de ordem** pela comparação da expansão em série de Taylor de $y(t)$ com a expansão em série de cada uma das funções $g_j(t,y)$. Por ser um precedimento extremamente trabalhoso, envolvendo muito algebrismo, não será apresentado neste texto, apresentando-se apenas os valores finais determinados.

- Método de Euler Simples Explícito $erro_{global} = \vartheta[h]$

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

- Método de Euler modificado $erro_{global} = \vartheta[h^2]$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

- Método de Euler aprimorado $erro_{global} = \vartheta[h^2]$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

- Método de Kutta $erro_{global} = \vartheta[h^3]$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

- Método de Runge-Kutta de quarta ordem padrão $erro_{global} = \vartheta[h^4]$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

- Método de Runge-Kutta-Gill de quarta ordem (Gill, 1951) $erro_{global} = \vartheta[h^4]$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/\sqrt{2}-1/2 & 1-1/\sqrt{2} & 0 & 0 \\ 1 & 0 & -1/\sqrt{2} & 1+1/\sqrt{2} & 0 \\ \hline & 1/6 & 1/3[1-1/\sqrt{2}] & 1/3[1+1/\sqrt{2}] & 1/6 \end{array}$$

- Método de Runge-Kutta de quinta ordem de Butcher (1987) $erro_{global} = \vartheta[h^5]$

0	0	0	0	0	0	0
1/4	1/4	0	0	0	0	0
1/4	1/8	1/8	0	0	0	0
1/2	0	$-1/2$	1	0	0	0
3/4	3/16	0	0	9/16	0	0
1	$-3/7$	2/7	12/7	$-12/7$	8/7	0
	7/90	0	32/90	12/90	32/90	7/90

- Método de Euler implícito de segunda ordem e um estágio $erro_{global} = \vartheta[h^2]$.

1/2	1/2
	1

Este método é também chamado de **método do ponto médio**, e pode ser expresso por

$$u_i = u_{i-1} + h f\left[t_{i-1} + \frac{h}{2}, \frac{u_i + u_{i-1}}{2}\right].$$

- Método de Runge-Kutta implícito de quarta ordem e dois estágios $erro_{global} = \vartheta[h^2]$

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	1/2	1/2

Para ilustrar essa notação, tomando como exemplo o método de Runge-Kutta explícito de quarta ordem na forma padrão, tem-se então:

$$\begin{aligned}
g_1 &= h f(t_{i-1}, u_{i-1}) \\
g_2 &= h f\left(t_{i-1} + \frac{h}{2}, u_{i-1} + \frac{1}{2} g_1\right) \\
g_3 &= h f\left(t_{i-1} + \frac{h}{2}, u_{i-1} + \frac{1}{2} g_2\right) \\
g_4 &= h f(t_{i-1} + h, u_{i-1} + g_3) \\
u_i &= u_{i-1} + \frac{1}{6}(g_1 + 2g_2 + 2g_3 + g_4).
\end{aligned}$$

7.4 Métodos de Integração de Passos Múltiplos

Voltando a apresentar o problema de valor inicial descrito na Equação 7.2:

$$\frac{dy(t)}{dt} = f[t, y(t)] \text{ para } t_{i-1} < t \leq t_i, \text{ sujeita à condição inicial: } y(t_{i-1}) = u_{i-1}.$$

Integrando ambos os lados dessa equação de t_{i-1} a t_i obtém-se:

$$y(t_i) = u_{i-1} + \int_{t_{i-1}}^{t_i} f[t, y(t)]dt = u_{i-1} + h \int_{z=0}^{z=1} f\left[t_{i-1} + zh, y(t_{i-1} + zh)\right] dz.$$

Considerando os valores da função $f[t, y(t)]$ no início dos m passos anteriores (em que o passo de integração h é constante):

t	$f[t,y(t)]$	$z=\dfrac{t-t_{i-1}}{h}$
t_{i-1}	f_{i-1}	0
t_{i-2}	f_{i-2}	-1
t_{i-3}	f_{i-3}	-2
$\vdots$	$\vdots$	$\vdots$
t_{i-m}	f_{i-m}	$-(m-1)$

e utilizando esses m valores de f na interpolação de Lagrange de grau $(m-1)$, tem-se:

$$f\left[t_{i-1}+zh,\, y(t_{i-1}+zh)\right] \approx p_{m-1}(z) = \sum_{j=1}^{m} \ell_j(z) f_{i-j} \text{ em que } \ell_j(z) = \prod_{k=1\neq j}^{m} \frac{z+(k-1)}{k-j}.$$

Empregando essa interpolação na integração de $f[t,y(t)]$, obtém-se:

$$y(t_i) \approx u_i = u_{i-1} + h\sum_{j=1}^{m} \beta_j^{(m)} f_{i-j} \text{ em que } \beta_j^{(m)} = \int_{z=0}^{z=1} \ell_j(z)dz.$$

Dando origem ao **Método de Adams[5]-Bashforth[6]**. Alguns valores dos coeficientes $\beta_j^{(m)}$ são a seguir tabelados.

j	1	2	3	4	5	6	*Erro acumulado*
$\beta_j^{(1)}$	1						$\vartheta[h]$
$2\beta_j^{(2)}$	3	-1					$\vartheta[h^2]$
$12\beta_j^{(3)}$	23	-16	5				$\vartheta[h^3]$
$24\beta_j^{(4)}$	55	-59	37	-9			$\vartheta[h^4]$
$720\beta_j^{(5)}$	1901	-2774	2616	-1274	251		$\vartheta[h^5]$
$1440\beta_j^{(6)}$	4277	-7923	9982	-7298	2877	-475	$\vartheta[h^6]$

Devido à natureza **explícita** do método de Adams-Bashforth o mesmo apresenta baixa estabilidade, para superar esse problema há uma modificação do método que inclui na integração de $f[t,y(t)]$ o valor de f no final do intervalo i, f_i, nesse caso, empregam-se os seguintes pontos nodais:

t	$f[t,y(t)]$	$z=\dfrac{t-t_{i-1}}{h}$
t_i	f_i	$+1$
t_{i-1}	f_{i-1}	0
t_{i-2}	f_{i-2}	-1
t_{i-3}	f_{i-3}	-2
$\vdots$	$\vdots$	$\vdots$
$t_{i-(m-1)}$	$f_{i-(m-1)}$	$-(m-2)$

Utilizando esses m valores de f na interpolação de Lagrange de grau $(m-1)$, tem-se:

$$f\left[t_{i-1}+zh,\, y(t_{i-1}+zh)\right] \approx \hat{p}_{m-1}(z) = \sum_{j=0}^{m-1} \hat{\ell}_j(z) f_{i-j} \text{ em que } \hat{\ell}_j(z) = \prod_{k=0\neq j}^{m-1} \frac{z+(k-1)}{k-j}.$$

Resultando no algoritmo **implícito**:

$$y(t_i) \approx u_i = u_{i-1} + h\sum_{j=0}^{m-1} \hat{\beta}_j^{(m)} f_{i-j} \text{ em que } \hat{\beta}_j^{(m)} = \int_{z=0}^{z=1} \hat{\ell}_j(z)dz.$$

[5]John Couch Adams (1819-1892).
[6]Francis Bashforth (1819-1912).

Dando origem ao **Método de Adams-Moulton**[7]. Alguns valores dos coeficientes $\hat{\beta}_j^{(m)}$ são a seguir tabelados.

j	1	2	3	4	5	6	*Erro acumulado*
$\hat{\beta}_j^{(1)}$	1						$\vartheta[h]$
$2\hat{\beta}_j^{(2)}$	1	1					$\vartheta[h^2]$
$12\hat{\beta}_j^{(3)}$	5	8	-1				$\vartheta[h^3]$
$24\hat{\beta}_j^{(4)}$	9	19	-5	1			$\vartheta[h^4]$
$720\hat{\beta}_j^{(5)}$	251	646	-264	106	-19		$\vartheta[h^5]$
$1440\hat{\beta}_j^{(6)}$	475	1427	-798	482	-173	27	$\vartheta[h^6]$

Observa-se que para $m=1$ o método de Adams-Bashforth recai no método de Euler explícito e o método de Adams-Moulton recai no método de Euler implícito.

Geralmente a implementação numérica desses algoritmos é feita em duas etapas:

1. **Etapa de Predição**: Método de Adams-Bashforth $u_i^{(0)} = u_{i-1} + h\sum_{j=1}^{m} \beta_j^{(m)} f_{i-j}$
2. **Etapa de Correção**: Método de Adams-Moulton $u_i^{(k+1)} = u_{i-1} + h\hat{\beta}_0^{(m)} f[t_i, u_i^{(k)}] + h\sum_{j=1}^{m-1} \hat{\beta}_j^{(m)} f_{i-j}$

   para $k = 0, 1, \cdots$.

Procedimentos desse tipo chamam-se de métodos **preditor-corretor** sendo de amplo emprego em códigos computacionais. Em cada passo de integração, a etapa de predição gera uma estimativa inicial para o método numérico, geralmente Newton-Raphson, de resolução da equação não linear da etapa de correção.

De uma forma geral os métodos de passos múltiplos podem ser descritos pela fórmula:

$$u_i = \sum_{j=1}^{k_1} a_{i,j} u_{i-j} + h_i \sum_{j=0}^{k_2} b_{i,j} f(t_{i-j}, u_{i-j}).$$

Admitindo-se nessa fórmula a variação do tamanho do passo de integração h_i com i.

Os chamados métodos de *retro-diferenciação* (***B**ackward **D**ifferentiation **F**ormula*: **BDF**) são os métodos de passos múltiplos em que $b_{i,j} = 0$ para $j > 0$, isto é,

$u_i = \sum_{j=1}^{k_1} a_{i,j} u_{i-j} + h_i b_{i,0} f(t_i, u_i)$. Simplificando a notação (fazendo $k_1 = m$ e para h constante): $\sum_{j=0}^{m} \alpha_j^{(m)} u_{i-j} = h_i f(t_i, u_i)$, em que $\alpha_j^{(m)} = -a_{i,j}/b_{i,0}$ com $a_{i,0} = -1$. O procedimento de determinação dos coeficientes $\alpha_j^{(m)}$ é apresentado a seguir para $m = 1, 2$ e 3.

- $m = 1 \Rightarrow y(z) \approx p_1(z) = z u_i + (1-z) u_{i-1}$ logo $p_1'(1) = u_i - u_{i-1} = hf[t_i, u_i]$.
- $m = 2 \Rightarrow y(z) \approx p_2(z) = \dfrac{z(z+1)}{2} u_i + (1-z^2) u_{i-1} + \dfrac{z(z-1)}{2} u_{i-2}$ logo
  $p_2'(1) = \dfrac{3}{2} u_i - 2u_{i-1} + \dfrac{1}{2} u_{i-2} = hf[t_i, u_i]$.
- $m = 3 \Rightarrow y(z) \approx p_3(z) = \dfrac{z(z+1)(z+2)}{6} u_i + \dfrac{(1-z^2)(z+2)}{2} u_{i-1} + \dfrac{z(z-1)(z+2)}{2} u_{i-2} + \dfrac{z(1-z^2)}{6} u_{i-3}$
  logo $p_3'(1) = \dfrac{11}{6} u_i - 3u_{i-1} + \dfrac{3}{2} u_{i-2} - \dfrac{1}{3} u_{i-3} = hf[t_i, u_i]$.

[7] Forest Ray Moulton (1872-1952).

Listando-se na tabela a seguir, os primeiros valores dos coeficientes $\alpha_j^{(m)}$.

j	0	1	2	3	4	5	6	*Erro acumulado*
$\alpha_j^{(1)}$	1	-1						$\vartheta[h]$
$2\alpha_j^{(2)}$	3	-4	1					$\vartheta[h^2]$
$6\alpha_j^{(3)}$	11	-18	9	-2				$\vartheta[h^3]$
$12\alpha_j^{(3)}$	25	-48	36	-16	3			$\vartheta[h^4]$
$60\alpha_j^{(4)}$	137	-300	300	-200	75	-12		$\vartheta[h^5]$
$60\alpha_j^{(5)}$	147	-360	450	-400	225	-72	10	$\vartheta[h^6]$

Uma das dificuldades de implementação de métodos de passos múltiplos é a necessidade do conhecimento de valores de u no início do passo em questão e em m passos anteriores. No início do processo de integração ($i = 1$), o valor de u é apenas conhecido no início do passo obrigando assim ao método de passos múltiplos ser o de menor ordem (método de Euler). Para superar esse problema, duas estratégias são adotadas: (i) começar os m primeiros passos de integração adotando um método de passo simples de ordem m (métodos de Runge-Kutta de ordem m, por exemplo); (ii) começar o processo com um método de um passo (Euler), no segundo passo adotar um método de dois passos ($m = 2$) e assim sucessivamente até atingir o número de passos desejado e, a partir desse ponto, segue com o método de m passos. A segunda estratégia é a mais utilizada em códigos computacionais correntes no mercado e a mesma é adotada com uma técnica adequada de seleção do tamanho de passo (que é variável ao longo do processo de integração).

7.5 O Conceito de Rigidez em Sistemas de EDOs

A estabilidade dos métodos explícitos de integração de EDO é garantida se o passo de integração for limitado por:

$$h \leq \frac{p}{|\Re(\lambda_{\text{máx}})|}$$

em que p é uma constante que depende do método empregado e $\lambda_{\text{máx}}$ é o valor característico do sistema que apresenta a parte real de maior valor em módulo. Por exemplo, o método de Euler simples explícito apresenta $p = 2$.

■ **Exemplo 7.6** Resolver os seguintes exemplos com o método de Euler explícito:

(a) $\dfrac{dy_1(t)}{dt} = -y_1(t)$, com $y_1(0) = 1{,}5$. A solução analítica é $y_1(t) = 1{,}5e^{-t} \longrightarrow \lambda_1 = -1$, implicando em $h \leq \dfrac{2}{1} = 2$. No caso de $t_f = 10$, são necessários no mínimo 5 passos.

(b) $\dfrac{dy_2(t)}{dt} = -1000y_2(t)$, com $y_2(0) = 0{,}5$. A solução analítica é $y_2(t) = 0{,}5e^{-1000t} \longrightarrow \lambda_1 = -1000$, implicando em $h \leq \dfrac{2}{1000} = 0{,}002$. No caso de $t_f = 10$, são necessários no mínimo 5000 passos.

(c) $\begin{pmatrix} \dfrac{dy_1(t)}{dt} \\ \dfrac{dy_2(t)}{dt} \end{pmatrix} = \begin{pmatrix} -500{,}5 & 499{,}5 \\ 499{,}5 & -500{,}5 \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}$ com $\begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, cuja solução analítica é $\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} 1{,}5e^{-t} + 0{,}5e^{-1000t} \\ 1{,}5e^{-t} - 0{,}5e^{-1000t} \end{pmatrix} \longrightarrow \lambda = \begin{pmatrix} -1 \\ -1000 \end{pmatrix}$, implicando no passo de integração $h \leq \dfrac{2}{1000} = 0{,}002$ imposto pela dinâmica mais rápida do sistema.

■

■ **Exemplo 7.7** Para ilustrar o conceito de rigidez, o seguinte exemplo é considerado: dois reatores químicos em série, onde é conduzida isotermicamente uma reação de primeira ordem, irreversível em fase líquida.

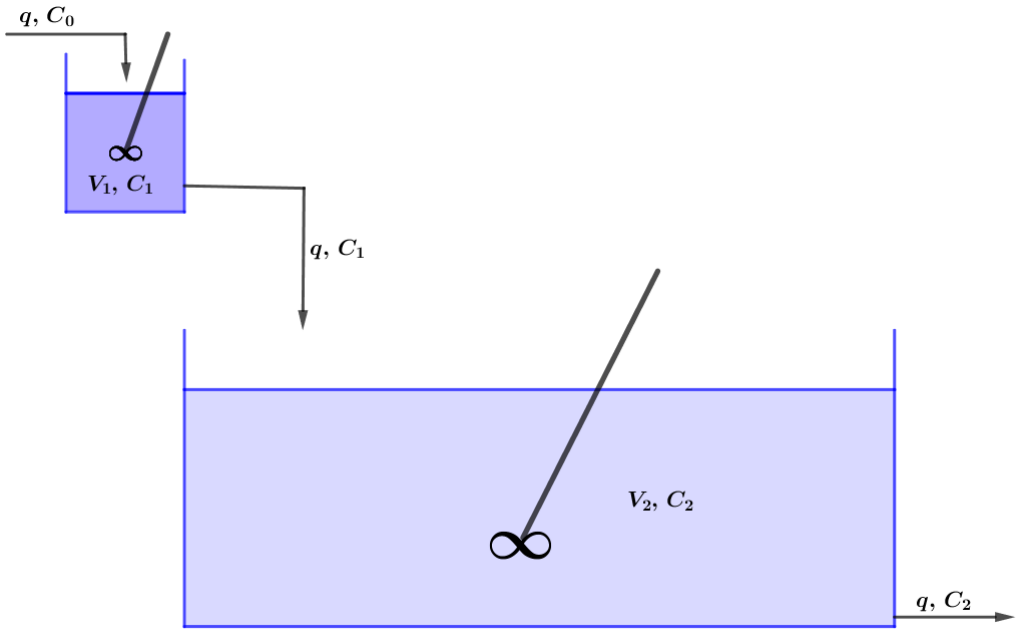


Figura 7.12: Dois reatores tanque contínuos em série.

Os balanços de massa do reagente em cada um dos reatores são dados por:

- Primeiro reator: $V_1\dfrac{dC_1(t)}{dt} = q[C_0(t) - C_1(t)] - kV_1C_1(t)$
- Segundo reator: $V_2\dfrac{dC_2(t)}{dt} = q[C_1(t) - C_2(t)] - kV_2C_2(t)$

Considerando que no início da contagem do tempo não ocorria reação alguma no interior dos reatores, tem-se: $C_1(0) = C_2(0) = 0$ e que exatamente em $t = 0$ o primeiro reator é alimentado por uma solução com concentração constante C_0.

Reescrevendo as equações de balanço em variáveis e parâmetros adimensionais:

$\tau = \dfrac{t}{V_1/q}$, $y_1(\tau) = \dfrac{C_1(t)}{C_0}$, $y_2(\tau) = \dfrac{C_2(t)}{C_0}$, $r = \dfrac{V_2}{V_1}$ e $Da = k\dfrac{V_1}{q}$, obtêm-se:

- Primeiro reator: $\dfrac{dy_1(\tau)}{d\tau} = [1 - y_1(\tau)] - Day_1(\tau)$ com $y_1(0) = 0$
- Segundo reator: $r\dfrac{dC_2(t)}{dt} = \dfrac{dy_1(\tau)}{d\tau} = [y_1(\tau) - y_2(\tau)] - rDay_2(\tau)$ com $y_2(0) = 0$

Considerando: $Da = 0,01$ e $r = 100$ (o segundo reator tem um volume 100 vezes maior que o primeiro), resulta:

- Primeiro reator: $\dfrac{dy_1(\tau)}{d\tau} = 1 - 1,01y_1(\tau)$ com $y_1(0) = 0$
- Segundo reator: $100\dfrac{dy_2(\tau)}{d\tau} = y_1 - 2y_2(\tau)$ com $y_2(0) = 0$

A solução analítica deste sistema de EDOs é $\mathbf{y}(\tau) = \begin{pmatrix} \dfrac{1-e^{-1,01\tau}}{1,01} \\ \dfrac{1-e^{-0,02\tau}}{2,02} + \dfrac{e^{-1,01\tau}-e^{-0,02\tau}}{99,99} \end{pmatrix}$.
Adotando as novas variáveis: $Y_1(\tau) = 1,01y_1(\tau)$ e $Y_2(\tau) = 2,02y_2(\tau)$ tem-se:

$$\mathbf{Y}(\tau) = \begin{pmatrix} 1-e^{-1,01\tau} \\ 1-e^{-0,02\tau} + \dfrac{e^{-1,01\tau}-e^{-0,02\tau}}{49,5} \end{pmatrix} \Rightarrow \lim_{\tau\longrightarrow\infty} \mathbf{Y}(\tau) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Note, na Figura 7.13, que a concentração de saída do primeiro reator varia, como era previsível, muito mais rápido do que a concentração de saída do segundo reator e, após o valor de $\tau = 5$ a concentração de saída do primeiro reator mantém-se praticamente constante e igual a seu valor estacionário final. Já a concentração de saída do segundo reator atinge o estado estacionário após $\tau = 200$. Essa diferença acentuada da velocidade de resposta das duas variáveis do problema é chamado de **rigidez** do sistema de EDOs, dito rígido.

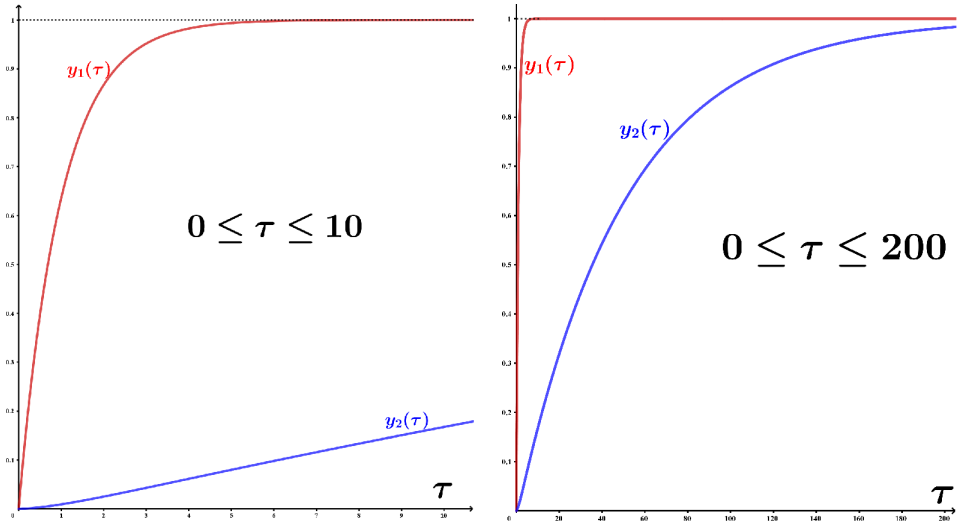


Figura 7.13: Partida de dois CSTRs.

■

A rigidez de um sistema de EDOs é caracterizada pela razão de rigidez (*stiffness ratio*: **SR**), que é a razão entre o módulo da parte real do valor característico que apresenta (em módulo) a maior parte real e o módulo da parte real do valor característico que apresenta (em módulo) a menor parte real.

Desse modo, no Exemplo 7.7 se tem: $\mathbf{SR} = \dfrac{1,01}{0,02} = 50,5 > 20$. Tipicamente problemas com $\mathbf{SR} < 20$ não são rígidos, para **SR** em torno de 1000 é considerado rígido e $\mathbf{SR} \geq 10^6$ é considerado muito rígido. Se o sistema é não linear a razão de rigidez é calculada com os valores característicos da matriz jacobiana do sistema.

Sob o ponto de vista numérico, a rigidez do sistema pode ser problemática, pois o passo de integração para os métodos explícitos deve satisfazer um critério relacionado ao módulo da parte real do maior valor característico do sistema, assim: $h \leq \dfrac{p}{|\Re(\lambda_{\text{máx}})|}$, em que $\lambda_{\text{máx}}$: é o valor de característico que apresenta a parte real de maior valor (em módulo). O tempo total de integração necessário para acompanhar toda a resposta dinâmica do sistema é, entretanto, escolhido de modo a

satisfazer um critério relacionado ao módulo da parte real do menor valor característico do sistema:

$$t_{total} = n_{total}h \geq \frac{5}{|\Re(\lambda_{min})|} \Rightarrow n_{total} > \frac{5}{p}\frac{|\Re(\lambda_{máx})|}{|\Re(\lambda_{min})|} = \frac{5}{p}\mathbf{SR}.$$

Podendo-se assim depreender que quanto maior for a razão de rigidez [**SR**] maior o número de passos de integração necessários e, em consequência, consumindo um grande tempo de computação. A alternativa para resolver problemas rígidos é utilizar algoritmos numéricos de integração que sejam implícitos, pois esses métodos são geralmente sempre estáveis não havendo restrições impostas à seleção do tamanho do passo de integração.

Uma maneira às vezes utilizada para contornar a rigidez do sistema é considerar a parte do sistema que tem a resposta mais rápida como se atingisse *instantaneamente* o estado estacionário final, essa simplificação é chamada de suposição de *estado quase-estacionário* (**QSSA**: *quasi steady-state assumption*) e é largamente empregada em Engenharia Química. No exemplo em questão isto equivaleria em considerar $Y_1(\tau) = 1$ e $Y_2(\tau) = 1 - e^{-0{,}02\tau}$ para $\tau > 0$. As curvas da Figura 7.14 representam o erro em $Y_2(\tau)$ quando se adota a suposição de *estado quase-estacionário* para a concentração de saída do primeiro reator.

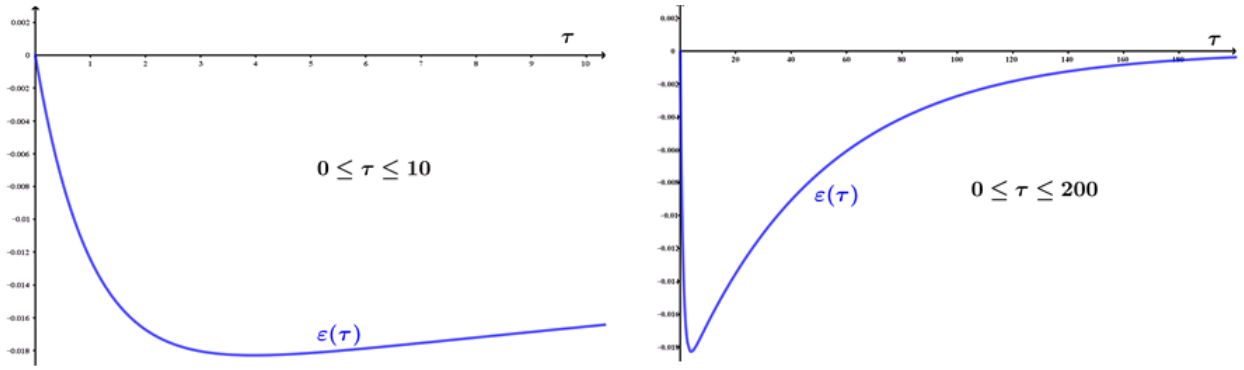


Figura 7.14: Erro em $Y_2(\tau)$ com a suposição de estado quase-estacionário para a concentração de saída do primeiro reator.

7.6 Restrições Algébricas e o Conceito de Índice Diferencial

Para ilustrar o conceito de restrições algébricas em sistemas de equações diferenciais ordinárias, considera-se um vaso de *flash* multicomponente, diagrama representado na Figura 7.15.

Adotando as seguintes hipóteses para a construção do modelo matemático:

- Mistura perfeita nas duas fases;
- Dinâmica da fase vapor desconsiderada;
- Entalpia da fase líquida: $h = C_p(T - T_{ref})$, $h_{ref} = 0$;
- Entalpia da fase vapor: $H = h + \lambda(T, P, y, x)$.

O modelo dinâmico deste processo é constituído pelas equações:

- Balanço de massa global: $\dfrac{dm(t)}{dt} = F(t) - V(t) - L(t)$
- Balanço de massa por componente:

$$\frac{d[m(t)x_i(t)]}{dt} = m(t)\frac{dx_i(t)}{dt} + x_i(t)\frac{dm(t)}{dt} = F(t)z_i(t) - V(t)y_i(t) - L(t)x_i(t)$$

$$m(t)\frac{dx_i(t)}{dt} = F(t)[z_i(t) - x_i(t)] - V(t)[y_i(t) - x_i(t)]$$

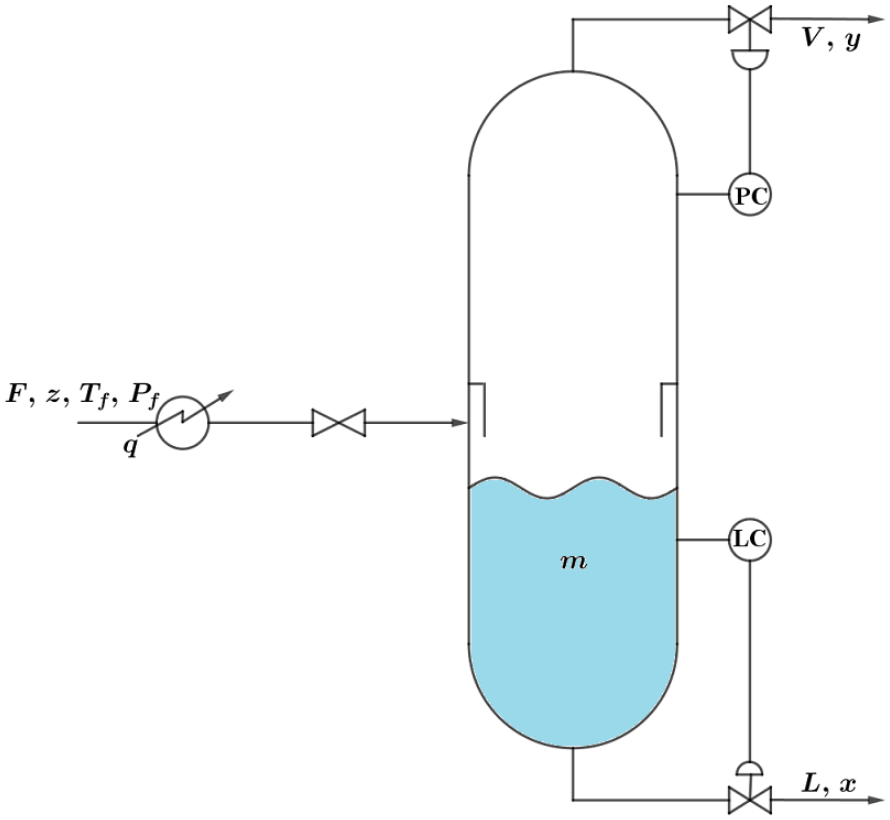


Figura 7.15: Vaso de *flash* multicomponente.

- Balanço de energia:
  $\frac{d[m(t)h(t)]}{dt} = m(t)\frac{dh(t)}{dt} + h(t)\frac{dm(t)}{dt} = F(t)h_f(t) - V(t)H(t) - L(t)h + q(t)$
  $m(t)C_p\frac{dT(t)}{dt} = F(t)C_p[T_f(t) - T(t)] - V(t)\lambda(t) + q(t)$
- Equilíbrio termodinâmico: $y_i(t) = K_i(t)x_i(t)$ em que $K_i(t) = f[T(t), P(t), \mathbf{x}(t), \mathbf{y}(t)]$
- Restrição das frações mássicas: $\sum_{i=1}^{n_c} x_i(t) = \sum_{i=1}^{n_c} y_i(t) = 1.$

Dando origem ao sistema de equações diferenciais ordinárias e equações algébricas, sistema de *Equações Algébrico-Diferenciais* (**EAD**):

$$\begin{cases} \frac{dm(t)}{dt} = F(t) - V(t) - L(t) \\ m(t)\frac{dx_i(t)}{dt} = F(t)[z_i(t) - x_i(t)] - V(t)[y_i(t) - x_i(t)] \text{ para } i = 1, 2, \cdots, n_c \\ m(t)C_p\frac{dT(t)}{dt} = F(t)C_p[T_f(t) - T(t)] - V(t)\lambda(t) + q(t) \\ \sum_{i=1}^{n_c} x_i(t)[1 - K_i(t)] = 0 \end{cases}.$$

Sujeito às condições iniciais: $\begin{cases} m(t_0) = m_0 \\ x_i(t_0) = x_{i,0} \text{ para } i = 1, 2, \cdots, n_c \\ T(t_0) = T_0 \end{cases}$,

em que: $\begin{cases} V(t) = f[P(t)] \text{ controlador de pressão} \\ L(t) = f[m(t)] \text{ controlador de nível} \\ K_i(t) = f[T(t), P(t), \mathbf{x}(t), \mathbf{y}(t)] \text{ para } i = 1, 2, \cdots, n_c \\ y_i(t) = K_i(t) x_i(t) \text{ para } i = 1, 2, \cdots, n_c \\ \lambda(t) = f[T(t), P(t), \mathbf{x}(t), \mathbf{y}(t)] \text{ calor latente de vaporização.} \end{cases}$

Este tipo de sistema de EADs pode ser representado genericamente por: $\mathbf{F}(t, \mathbf{v}, \dot{\mathbf{v}}, \mathbf{w}, \mathbf{u}) = \mathbf{0}$, em que $\mathbf{v}$ é o vetor das variáveis diferenciais, $\mathbf{w}$ é o vetor das variáveis algébricas e $\mathbf{u}$ é o vetor das variáveis de entrada.

Frequentemente as equações algébricas são resolvidas em um processo iterativo embutido no método de integração das equações diferenciais. Entretanto, este tipo de procedimento é muito mais lento do que o procedimento de resolução simultânea das equações diferenciais e algébricas. Grande cautela deve haver na *inicialização consistente* do sistema, que exige a satisfação das restrições algébricas em $t = t_0$.

Métodos numéricos de resolução de sistemas de EADs: tais procedimentos transformam o sistema de EADs em um sistema puramente algébrico seja pela substituição de $\mathbf{v}'(t)$ (métodos de passos múltiplos, BDF), seja pela substituição de $\mathbf{v}(t)$ (métodos de passo simples, tipo métodos de Runge-Kutta), por uma aproximação numérica:

- Métodos de passos múltiplos: $\mathbf{v}'(t) \approx \mathbf{G}[\mathbf{v}(t)] \Rightarrow \breve{\mathbf{F}}(t, \mathbf{v}, \mathbf{w}, \mathbf{u}) = \mathbf{0}$;
- Métodos de passo simples: $\mathbf{v}(t) \approx \mathbf{H}[\mathbf{v}'(t)] \Rightarrow \hat{\mathbf{F}}(t, \mathbf{v}', \mathbf{w}, \mathbf{u}) = \mathbf{0}$.

Tais sistemas algébricos são geralmente resolvidos pelo método de Newton-Raphson e suas modificações, conforme esquematizado na Figura 7.16.

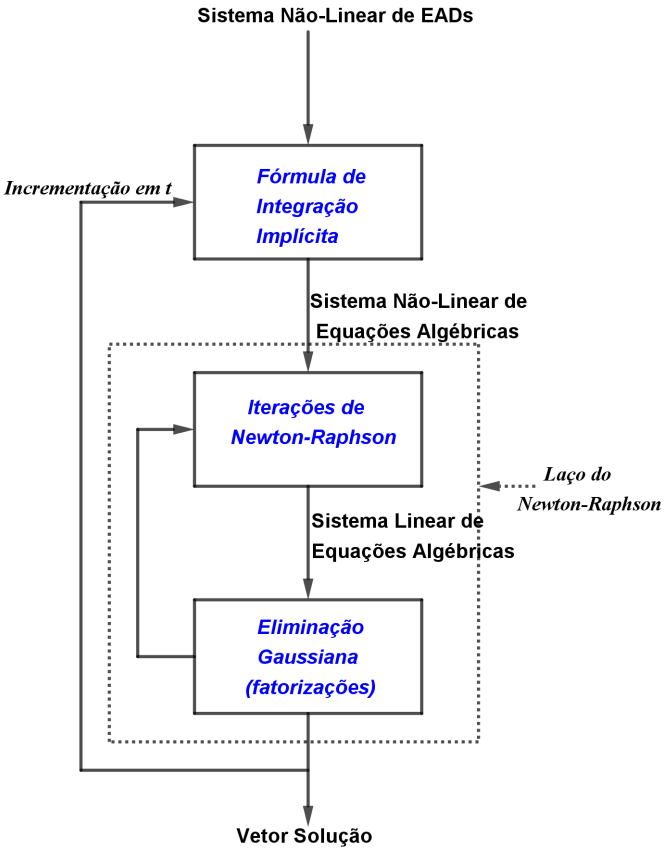


Figura 7.16: Procedimento de solução de sistemas de EADs.

7.6.1 Problemas de Índice em Sistemas de Equações Algébrico-Diferenciais

Para ilustrar o problema de índice em equações algébrico-diferenciais o problema clássico da dinâmica do pêndulo simples, Figura 7.17, é analisado.

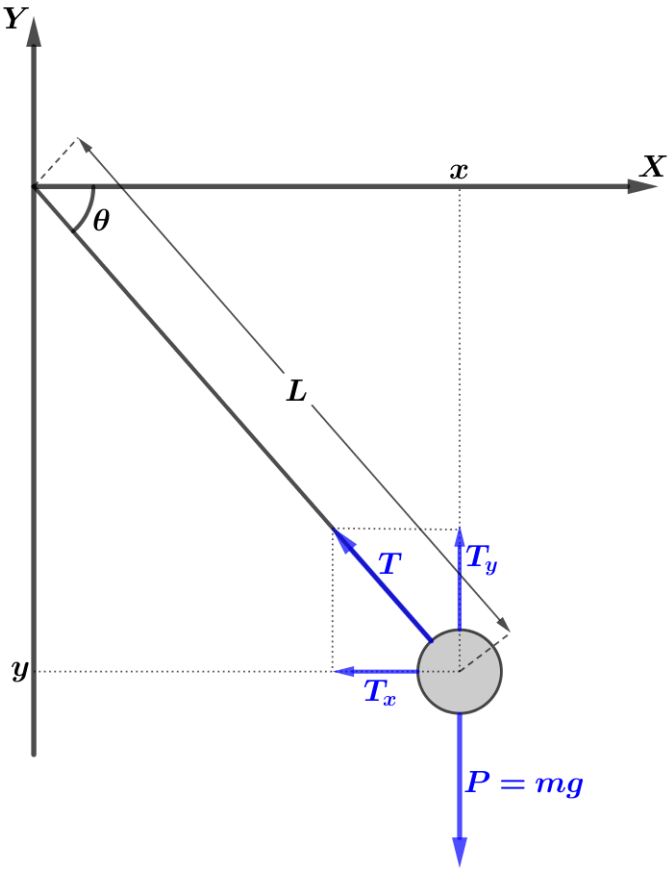


Figura 7.17: O pêndulo simples.

O modelo dinâmico deste processo é constituído pelas equações:

- Equações do movimento na direção x: $\begin{cases} \dfrac{d\mathrm{x}(t)}{dt} = \mathrm{v}_x(t) \\ m\dfrac{d\mathrm{v}_x(t)}{dt} = -T_x(t) = -T(t)\cos(\theta) \end{cases}$
- Equações do movimento na direção y: $\begin{cases} \dfrac{d\mathrm{y}(t)}{dt} = \mathrm{v}_y(t) \\ m\dfrac{d\mathrm{v}_y(t)}{dt} = T_y(t) - mg = T(t)\,\mathrm{sen}(\theta) - mg \end{cases}$

Em vista de $\cos(\theta) = \dfrac{\mathrm{x}}{L}$ e $\mathrm{sen}(\theta) = -\dfrac{\mathrm{y}}{L}$ obtêm-se:

- Equações do movimento na direção x: $\begin{cases} \dfrac{d\mathrm{x}(t)}{dt} = \mathrm{v}_x(t) \\ m\dfrac{d\mathrm{v}_x(t)}{dt} = -T(t)\dfrac{\mathrm{x}}{L} \end{cases}$
- Equações do movimento na direção y: $\begin{cases} \dfrac{d\mathrm{y}(t)}{dt} = \mathrm{v}_y(t) \\ m\dfrac{d\mathrm{v}_y(t)}{dt} = -T(t)\dfrac{\mathrm{y}}{L} - mg \end{cases}$

Além dessas equações, as variáveis dependentes $\mathrm{x}(t)$ e $\mathrm{y}(t)$ devem satisfazer à restrição algébrica $\mathrm{x}^2(t) + \mathrm{y}^2(t) = L^2$.

Condições iniciais: $\mathrm{x}(0)=\mathrm{x}_0$, $\mathrm{y}(0)=\mathrm{y}_0=-\sqrt{L^2-\mathrm{x}_0^2}$, $\mathrm{v}_x(0)=\mathrm{v}_y(0)=0$ (o pêndulo parte do *repouso*). Verifica-se que sem a manipulação das equações não se pode determinar o valor de $T(t)$.

Para facilitar o manuseio das equações do modelo adotam-se as variáveis adimensionais: $x=\dfrac{\mathrm{x}}{L}$, $y=\dfrac{\mathrm{y}}{L}$, $v_x=\dfrac{\mathrm{v}_x}{\sqrt{Lg}}$, $v_y=\dfrac{\mathrm{v}_y}{\sqrt{Lg}}$, $\lambda=\dfrac{T}{mg}$ e $\tau=t\sqrt{\dfrac{g}{L}}$, dando origem à equações adimensionais:

- Equações do movimento na direção x: $\begin{cases}\dfrac{dx(t)}{d\tau}=v_x(\tau)\\ \dfrac{dv_x(t)}{d\tau}=-\lambda(\tau)x(\tau)\end{cases}$
- Equações do movimento na direção y: $\begin{cases}\dfrac{dy(t)}{d\tau}=v_y(\tau)\\ \dfrac{dv_y(t)}{d\tau}=-\lambda(\tau)y(\tau)-1\end{cases}$
- Restrição algébrica: $x^2(\tau)+y^2(\tau)=1$.

Condições iniciais: $x(0)=x_0\,(|x_0|\leq 1)$, $y(0)=y_0=-\sqrt{1-x_0^2}$, $v_x(0)=v_y(0)=0$ (o pêndulo parte do *repouso*). Verifica-se que ainda não se pode determinar o valor de $\lambda(\tau)$.

Diferenciando em relação à τ a restrição algébrica:

$x^2(\tau)+y^2(\tau)=1\Rightarrow 2[v_x(\tau)x(\tau)+v_y(\tau)y(\tau)]=0$ dando origem a outra restrição algébrica (**restrição escondida**) $v_x(\tau)x(\tau)+v_y(\tau)y(\tau)=0$ em que ainda não aparece a variável $\lambda(\tau)$. Diferenciando em relação à τ esta nova restrição algébrica, resulta:

$\dfrac{dv_x(\tau)}{d\tau}x(\tau)+[v_x(\tau)]^2+\dfrac{dv_y(\tau)}{d\tau}y(\tau)+[v_y(\tau)]^2=0$ e substituindo nesta expressão as equações de $\dfrac{dv_x(\tau)}{d\tau}$ e $\dfrac{dv_x(\tau)}{d\tau}$, obtém-se: $-\lambda(\tau)[x^2(\tau)+y^2(\tau)]+[v_x(\tau)]^2+[v_x(\tau)]^2-y(\tau)=0$. Em vista de $x^2(\tau)+y^2(\tau)=1$, chega-se finalmente a uma segunda restrição escondida:

$$\lambda(\tau)=[v_x(\tau)]^2+[v_y(\tau)]^2-y(\tau).$$

Para obter a derivada da variável $\lambda(\tau)$ deve-se diferenciar novamente esta última equação dando origem a $\dfrac{d\lambda(\tau)}{d\tau}=-2\lambda(\tau)[v_x(\tau)x(\tau)+v_y(\tau)y(\tau)]-3v_y(\tau)=-3v_y(\tau)$. Chegando-se ao sistema constituído apenas por equações diferenciais ordinárias:

$$\begin{cases}\dfrac{dx(t)}{d\tau}=v_x(\tau)\\ \dfrac{dv_x(t)}{d\tau}=-\lambda(\tau)x(\tau)\\ \dfrac{dy(t)}{d\tau}=v_y(\tau)\\ \dfrac{dv_y(t)}{d\tau}=-\lambda(\tau)y(\tau)-1\\ \dfrac{d\lambda(\tau)}{d\tau}=-3v_y(\tau)\end{cases}.$$

Entretanto, para se ter condições inciais consistentes as restrições algébricas devem ser todas satisfeitas inclusive para $\tau=0$, isto é:

$$\begin{cases}x^2(\tau)+y^2(\tau)=1\\ v_x(\tau)x(\tau)+v_y(\tau)y(\tau)=0\\ \lambda(\tau)=[v_x(\tau)]^2+[v_y(\tau)]^2-y(\tau)\end{cases}\Rightarrow\begin{cases}x^2(0)+y^2(0)=1\\ v_x(0)x(0)+v_y(0)y(0)=0\\ \lambda(0)=[v_x(0)]^2+[v_y(0)]^2-y(0)\end{cases}.$$

Por exemplo, *soltando* do repouso o pêndulo na posição horizontal positiva, isto é $x(0) = x_0 = 1$ e $v_y(0) = v_{y_0} = 0$, as demais variáveis dependentes devem ser $y_0 = 0$, $v_{x_0} = 0$ e $\lambda_0 = 0$.

A necessidade de diferenciação do sistema original para resgatar as expressões de todas as variáveis dependentes dá origem ao conceito de **índice diferencial** de um sistema de EADs.

Definição 7.6.1 — Índice diferencial, ν. Considerando a forma geral de um sistema de EADs: $\mathbf{F}(t, \mathbf{v}, \mathbf{v}', \mathbf{u}) = \mathbf{0}$, em que $\mathbf{v} \wedge \mathbf{v}' \in \Re^n$ é o vetor das variáveis de estado e de suas derivadas temporais, respectivamente, e $\mathbf{u} \in \Re^r$ é o vetor das variáveis de entrada. O índice diferencial deste sistema é o número mínimo de vezes que todo ou parte do sistema deve ser diferenciado em relação à variável independente t para determinar $\mathbf{v}'(t)$ como uma função contínua de t e $\mathbf{v}(t)$.

De acordo com esta definição, o problema do pêndulo simples é um sistema de EADs com índice diferencial igual a 3. O procedimento de diferenciações sucessivas do problema do pêndulo para chegar a essa conclusão é conhecido como processo de *redução de índice*. Sistemas de EADs com $\nu > 1$ são ditos de *índice elevado* e que podem apresentar problemas de índice (Brenan, Campbell e Petzold, 1996). Observe que um sistema com $\nu = 0$ é na verdade um sistema de EDOs.

Um outro conceito importante para sistemas de EADs é o número de **graus de liberdade dinâmicos** (GLD), ou seja, o número de condições iniciais arbitrárias que o sistema admite, ou também o número verdadeiro de estados do sistema. O GLD equivale a $2n - n_e$, em que n é a dimensão do sistema de EADs original e n_e é a dimensão do **sistema estendido**, isto é, do sistema original aumentado pelas equações resultantes do processo de redução de índice. O GLD também pode ser determinado pelo número de *variáveis diferenciais* (aquelas em que suas derivadas temporais aparecem explicitamente no sistema de EADs original, caso contrário são chamadas de *variáveis algébricas*) menos o número de restrições escondidas.

No exemplo do pêndulo, $n = 5$ e $n_e = 8$, possui 4 variáveis diferenciais e duas restrições escondidas, logo esse problema possui apenas $2 \times 5 - 8 = 4 - 2 = 2$ graus de liberdade dinâmicos, admitindo portanto somente duas condições iniciais arbitrárias, que no exemplo foram $x(0) = x_0 = 1$ e $v_y(0) = v_{y_0} = 0$.

7.7 Problemas Propostos

Problema 7.1 Um balão de destilação aberto à atmosfera contém uma mistura binária com massa total m_0 e composição conhecida em $t = 0$. Exatamente em $t = 0_+$ a solução passa a destilar, com a composição da fase vapor em equilíbrio com a composição da fase líquida, expressa pela relação de equilíbrio termodinâmico: $y_1 = f(x_1)$, sendo "1" o componente mais volátil. Deseja-se saber a composição da mistura líquida no balão no instante em que contiver uma massa total conhecida $m_{final} < m_0$.

As equações diferenciais que *governam* este processo são:

1. Balanço global:
$$\frac{dm(t)}{dt} = -D(t) \text{ com } m(0) = m_0$$
2. Balanço do componente 1 :
$$\frac{d[m(t)x_1(t)]}{dt} = -D(t)y_1(t) = -D(t)f[x_1(t)] \, com x_1(0) = x_{1_0}$$

Adotando como nova variável independente a variável $m(t)$ (que é monótona decrescente) e aplicando a *regra da cadeia* obtém-se:

$$\frac{dm(t)x_1(t)}{dt} = \frac{dm(t)}{dt}\frac{d[mx_1(m)]}{dm} = -D(t)\left[x_1(m) + m\frac{dx_1(m)}{dm}\right] = -D(t)f[x_1(m)],$$

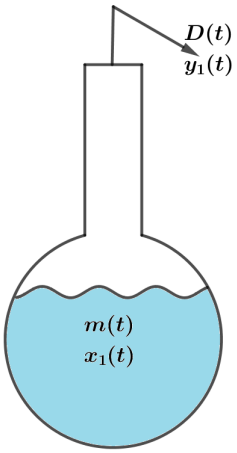


Figura 7.18: Balão de destilação de mistura binária.

ou seja:

$$\frac{dx_1(m)}{dm} = \frac{f[x_1(m)] - x_1(m)}{m} \text{ com } x_1(m_0) = x_{1_0}, \text{ integração de } m_0 \text{ para } m_{final} < m_0.$$

Adotando como nova variável independente a variável adimensional $\tau = \dfrac{m_0 - m(t)}{m_0 - m_{final}} \Rightarrow$

$$\begin{cases} m(t) = m_0 \longrightarrow \tau = 0 \\ m(t) = m_{final} \longrightarrow \tau = 1 \end{cases} \quad \text{e} \quad \frac{d\tau}{dm} = -\frac{1}{m_0 - m_{final}}.$$

A nova equação de balanço do componente 1 é então: $-\dfrac{1}{m_0 - m_{final}} \dfrac{dx_1(\tau)}{d\tau} = \dfrac{f[x_1(\tau) - x_1(\tau)]}{m}$.

Reagrupando os termos e definindo o parâmetro $\alpha = \dfrac{m_0}{m_0 - m_{final}} > 1$, chega-se a:

$$\frac{dx_1(\tau)}{d\tau} = -\left(\frac{f[x_1(\tau)] - x_1(\tau)}{\alpha - \tau}\right) \text{ com} x_1(0) = x_{1_0}, \text{ integração de } \tau = 0 \text{ a } \tau = 1.$$

Particularizando o problema para a destilação de uma mistura binária de n-octano e n-heptano (componente mais volátil 1) conduzida à pressão atmosférica. Sabendo-se que no início da batelada o balão contém $25\,mol$ de n-heptano e $75\,mol$ de n-octano ($m_0 = 100\,mol$), que no tempo final o balão contém $10\,mol$ da mistura ($m_{final} = 10\,mol$) e que à pressão atmosférica a relação de equilíbrio entre a composição molar do n-heptano na fase líquida e na fase vapor é dada por:

$y_1(x_1) = \dfrac{2,16x_1}{1 + 1,16x_1}$, (o que equivale a considerar a volatilidade relativa da mistura igual a 2,16).

Represente a variação da composição da mistura ao longo da destilação.

Problema 7.2 Considere o balão de destilação do Problema 7.1 com um condensador na saída do vapor, em acordo com a Figura 7.19.

As equações diferenciais que *governam* este processo são:

1. Balanço global:

$$\frac{dm(t)}{dt} = -D(t) \text{ com } m(0) = m_0$$

2. Balanço do n-heptano no balão:

$$\frac{d[m(t)x_1(t)]}{dt} = -(1+R)D(t)y_1(t) + RD(t)x_{c_1}(t) \text{ com } x_1(0) = x_{1_0}$$

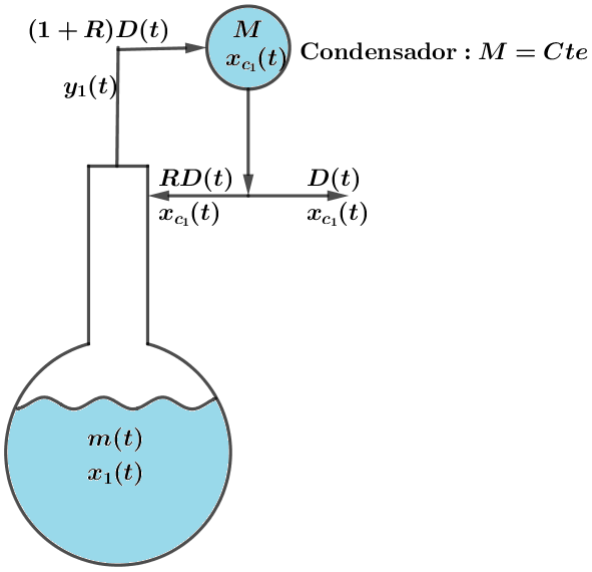


Figura 7.19: Balão de destilação de mistura binária com condensador.

3. Balanço do n-heptano no condensador:

$$M\frac{d[x_{c_1}(t)]}{dt} = (1+R)D(t)[y_1(t) - x_{c_1}(t)] \text{ com } x_{c_1}(0) = x_{c_{1_0}}.$$

A relação de equilíbrio entre a composição molar do n-heptano na fase líquida e na fase vapor é dada por: $y_1(x_1) = y_{eq}(x_1) = \dfrac{2,16x_1}{1+1,16x_1}$.

Reescreva as equações de balanço em termos da nova variável independente $\tau = \dfrac{m_0 - m(t)}{m_0 - m_{final}}$.

Determine $x_1(\tau)$ e $x_{c_1}(\tau)$ utilizando os seguintes valores numéricos:

$x_{1_0} = 0,75; x_{c_{1_0}} = y_{eq}(x_{1_0}) = 0,86631016; m_0 = 100 mold;$,$M = 10 mol$ e $m_{final} = 10 mol$.

Problema 7.3 Considere o seguinte modelo cinético de reação: $A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B \xrightarrow{k_3} C$ conduzida em batelada em um reator de mistura, iniciando-se com o componente A puro. A variação da concentração de A e de B com o tempo, na forma adimensional, é descrita pelo sistema de EDOs:

$$\begin{cases} \dfrac{dx_1(\tau)}{d\tau} = -\left(1+\dfrac{k_1}{k_2}\right)x_1(\tau) + x_2(\tau) \text{ com } x_1(0) = 1 \\ \dfrac{dx_2(\tau)}{d\tau} = \dfrac{k_1}{k_2}x_1(\tau) - \left(1+\dfrac{k_3}{k_2}\right)x_2(\tau) \text{ com } x_2(0) = 0 \end{cases}.$$

Sendo: $\dfrac{k_1}{k_2} = 10000$ e $\dfrac{k_3}{k_2} = 4$.

Determine $x_1(\tau)$ e $x_2(\tau)$ pelo método de Euler implícito com intervalo de integração constante. Indique claramente o passo de integração adequado e o número total de passos necessários para se acompanhar todo o processo dinâmico, isto é, o tempo necessário para $x_1(\tau) \longrightarrow 0$ e $x_2(\tau) \longrightarrow 0$..

Problema 7.4 A simulação da partida de um reator químico pode ser obtida através da resolução de equação diferencial ordinária (em forma adimensional):

$$\frac{dx(t)}{dt} = 1 - x(t) - \frac{7[x(t)]^2}{4+8x(t)} \text{ com } x(0) = 0.$$

Propõe-se resolver numericamente esta equação pelo **método do ponto médio** (Runge-Kutta de segunda ordem, implícito com um estágio), armazenando os valores de $x(t)$ em $t = 1, 2, 3$ e 4. Determine estes valores usando com passos de integração $h = 0,5$ e $0,25$ e refine estes resultados baseado na ordem de acurácia do método. Mostre claramente o procedimento recursivo empregado.

Problema 7.5 Em um sistema fechado com três componentes o seguinte esquema cinético (modelo cinético de Robertson, 1966) ocorre:

$$\begin{cases} A \xrightarrow{k_1} B \\ B + C \xrightarrow{k_2} A + C \\ 2B \xrightarrow{k_3} B + C \end{cases} .$$

A variação da concentração de A, B e C com o tempo, na forma adimensional, é descrita pelo sistema de EDOs:

$$\begin{cases} \dfrac{dx_1(t)}{dt} = k_2 x_2(t) x_3(t) - k_1 [x_1(t)]^2 \text{ com } x_1(0) = 1 \\ \dfrac{dx_2(t)}{dt} = k_1 [x_1(t)]^2 - k_2 x_2(t) x_3(t) - k_3 [x_2(t)]^2 \text{ com } x_2(0) = 0 \\ \dfrac{dx_3(t)}{dt} = k_3 [x_2(t)]^2 \text{ com } x_3(0) = 0 \end{cases} ,$$

sendo: $x_1(t) = \dfrac{C_A(t)}{C_{A_0}}, x_2(t) = \dfrac{C_B(t)}{C_{A_0}}$ e $x_3(t) = \dfrac{C_C(t)}{C_{A_0}}$.

Calcule a variação de $x_1(t), x_2(t)$ e $x_3(t)$ com t utilizando os seguintes valores das constantes cinéticas $k_1 = 0,08\text{s}^{-1}, k_2 = 2,00 \times 10^4 \text{s}^{-1}$ e $k_3 = 6,00 \times 10^7 \text{s}^{-1}$. Note que para todo t tem-se $x_1(t) + x_2(t) + x_3(t) = 1$.

Observação: A integração das equações é bastante facilitada adotando-se a consideração de estado quase-estacionário para a variável dependente $x_2(t)$.

Problema 7.6 Dinâmica de populações com interação . As equações diferenciais abaixo descrevem o comportamento dinâmico da sobrevivência de duas espécies animais em um mesmo habitat. A variável N_1 representa o número de elementos da espécie 1 que é um ser herbívoro para o qual há abundância de alimentos, a variável N_2 representa o número de elementos da espécie 2 que é o ser predador que se alimenta da espécie 1. As equações diferenciais que descrevem esta dinâmica é: (sendo a unidade da variável independente $t = ano$)

$$\begin{cases} \dfrac{dN_1(t)}{dt} = \alpha N_1(t) - \beta N_1(t) N_2(t) \text{ com } N_1(0) = N_{1_0} \\ \dfrac{dN_2(t)}{dt} = -\gamma N_2(t) + \delta N_1(t) N_2(t) \text{ com } N_2(0) = N_{2_0} \end{cases}$$

Com os valores de $\alpha = \dfrac{2}{5}, \beta = \dfrac{1}{90}, \gamma = \dfrac{1}{5}$ e $\delta = \dfrac{1}{3500}$ e as condições iniciais $N_{1_0} = 60$ e $N_{2_0} = 10$. Determine, através da integração numéricas das equações diferenciais do sistema, a variação temporal das variáveis N_1 e N_2 de $t = 0$ (início da contagem do tempo) até $t = 50\,anos$. Baseado nos resultados obtidos, estime a periodicidade da resposta dinâmica do sistema.

Problema 7.7 A partida de um reator, no qual há uma válvula instalada na saída, é descrito (em forma adimensional) pelas equações diferenciais ordinárias:

$$\begin{cases} \dfrac{dh(t)}{dt} = 1 - \sqrt{h(t)} \text{ com } h(0) = 0 \\ \dfrac{dc(t)}{dt} = \dfrac{1 - c(t) - Da\, h(t) [c(t)]^2}{h(t)} \text{ com } c(0) = 1 \end{cases} ,$$

sendo $h(t)$ a altura de líquido no reator, $c(t)$ a concentração do reagente no interior do reator (ambas variáveis adimensionais) no tempo t (também em forma adimensional).

Mostre como abordar a aparente singularidade da equação de balanço de reagente em $t = 0$. Após a regularização desta singularidade, determine a variação temporal de $c(t)$ até atingir sua concentração de operação (estado estacionário final) considerando o parâmetro Da (número de Damköhler) $= 2$.

Problema 7.8 Um reator tubular conduz adiabaticamente uma reação em fase gasosa exotérmica e irreversível, as equações que descrevem as variações de concentração de reagente e da temperatura ao longo do reator são, em forma adimensional e considerando o escoamento pistonado (*Plug Flow Reactor*):

$$\begin{cases} \text{Balanço do Reagente: } \dfrac{dy(x)}{dx} = -Da\, y(x) \exp\left[\gamma\left(1 - \dfrac{1}{\theta(x)}\right)\right] \text{ com } y(0) = 1 \\ \text{Balanço de Energia: } \dfrac{d\theta(x)}{dx} = \beta\, Da\, y(x) \exp\left[\gamma\left(1 - \dfrac{1}{\theta(x)}\right)\right] \text{ com } \theta(0) = 1 \end{cases}.$$

Multiplicando a primeira equação por β (fator de exotermicidade) e adicionando o resultando ao balanço de energia, obtém-se:

$$\frac{d\theta(x)}{dx} + \beta\frac{dy(x)}{dx} = \frac{d[\theta(x) + \beta y(x)]}{dx} = 0 \Rightarrow \theta(x) + \beta y(x) = C^{te} = \theta(0) + \beta y(0) = 1 + \beta.$$

Permitindo expressar $\theta(x) = 1 + \beta[1 - y(x)]$, que substituído no balanço do reagente dá origem a:

$$\text{Balanço do Reagente: } \frac{dy(x)}{dx} = -Da\, y(x) \exp\left(\frac{\gamma\beta[1 - y(x)]}{1 + \beta[1 - y(x)]}\right) \text{ com } y(0) = 1.$$

As variáveis e parâmetros adimensionais do problema são:

$$x = \frac{z}{L}, y = \frac{C}{C_0}, \theta = \frac{T}{T_0}, Da = \frac{k_0 L}{v_z}, \gamma = \frac{E}{RT_0} \text{ e } \beta = \frac{C_0(-\Delta H)}{\rho c_P T_0}.$$

Utilizando os dados: $L = 2m$; $C_0 = 0,03\, kmol/m^3$; $T_0 = 700\, K$; $(-\Delta H) = 10^4\, kJ/kmol$; $c_p = 1,0\, kJ/(kg\, K)$; $E = 100\, kJ/kmol$; $\rho = 1,2\, kg/m^3$; $k_0 = 5z, s^{-1}$ e $R = 8,31451\, kJ/kmol/K$ (constante universal dos gases), determine os perfis de y e θ ao longo de x.

Problema 7.9 Uma reação química irreversível é conduzida em um reator tanque de mistura de forma semi-contínua (*batelada alimentada*). A variação da concentração do reagente com o tempo é obtida pela resolução da equação de balanço:

(a) Fase de alimentação: $0 < t \leq \dfrac{V_{máx}}{q}$

$$V(t)\frac{dC(t)}{dt} = q[C_0 - C((t))] - \left(\frac{kC(t)}{1 + \alpha C(t)}\right) V(t) \text{ com } C(0) = C_0.$$

$$\text{Sendo;} \begin{cases} V(t) = qt \text{ volume da mistura no reator} \\ C(t) \text{ concentração do reagente no interior do reator} \\ q \text{ vazão volumétrica de alimentação (constante)} \\ C_0 \text{ concentração do reagente na alimentação do reator (constante)} \\ k \text{ e } \alpha \text{ constantes cinéticas} \\ V_{máx} \text{ volume máximo do reator} \end{cases}.$$

(b) Fase da batelada propriamente dita: $t > \dfrac{V_{máx}}{q}$

$$\frac{dC(t)}{dt} = -\left(\frac{kC(t)}{1 + \alpha C(t)}\right).$$

Devido à singularidade em $t = 0,$ pois $V(0) = 0,$ métodos de integração explícitos não podem ser aplicados diretamente, mostre como remover esta singularidade. Resolva a seguir o problema, escolhendo as constantes, os parâmetros e condição inicial pertinentes.

Problema 7.10 Em um biorreator contínuo de tanque agitado, a volume constante $V = 10m^3$ e temperatura fixa de $35\,^{\circ}C$, com alimentação de substrato a uma concentração de $x_{2f} = 4,0kg/m^3$ e vazão de alimentação $F = 3m^3/h$, ocorre uma reação de fermentação descrita pelo seguinte modelo:

$$\begin{cases} \dfrac{dx_1(t)}{dt} = \dfrac{\mu_{máx}x_1(t)x_2(t)}{k_m + x_2(t)} - \left(\dfrac{F}{V} + k_d\right)x_1(t) \text{ com } x_1(0) = 1,65 \\ \dfrac{dx_2(t)}{dt} = \dfrac{F}{V}[x_{2f} - x_2(t)] - \dfrac{1}{Y}\left(\dfrac{\mu_{máx}x_1(t)x_2(t)}{k_m + x_2(t)}\right) \text{ com } x_2(0) = 0,20 \end{cases},$$

em que: $x_1(t)$ é a concentração de células (biomassa), $x_2(t)$ é a concentração de substrato, ambas em kg/m^3, $\mu_{máx} = 0,53\,h^{-1}$ é a taxa máxima de crescimento específico, $k_d = 0,01\,h^{-1}$ é a constante de taxa de morte celular, $k_m = 0,12\,kg/m^3$ é a constante de saturação da taxa de crescimento celular e $Y = 0,4$ é o fator de rendimento da fermentação.

Resolva o problema aplicando o método simples de Euler explícito, selecionando o valor do passo de integração que assegure uma solução numérica estável e não-oscilatório. Refaça o problema pelo método simples de Euler implícito e pelo método do ponto médio, demonstrando em ambos os casos a estabilidade dos métodos. Verifique em todos seus resultados a capacidade de prever o valor do estado estacionário final do processo igual a $\mathbf{x}_{final} = \begin{pmatrix} 1,48293255 \\ 0,16909091 \end{pmatrix}$.

8. Introdução à Otimização

No contexto de otimização, os problemas são tratados usando as seguintes definições:
função objetivo: é a função matemática cujo máximo ou mínimo deseja-se determinar.
variáveis de decisão: são as variáveis independentes do problema de otimização. Correspondem, em número, ao excesso de variáveis em relação ao número de equações (restrições de igualdade), isto é, o número de **graus de liberdade** do sistema.
restrições: são os limites impostos ao sistema ou estabelecidos pelas leis naturais que governam o comportamento do sistema, a que estão sujeitas as variáveis de decisão. As restrições podem ser de igualdade (equações) ou de desigualdade (inequações).
região de busca: ou região viável, é a região do espaço definido pelas variáveis de decisão, delimitada pelas restrições, em cujo interior ou em cuja fronteira se localiza o ótimo da função objetivo.

■ **Exemplo 8.1 — Formulação do Problema.** No processo de extração por solvente puro, ilustrado na Figura 8.1, deseja-se encontrar a condição de operação com a maior lucratividade possível (Perlingeiro, 2005).

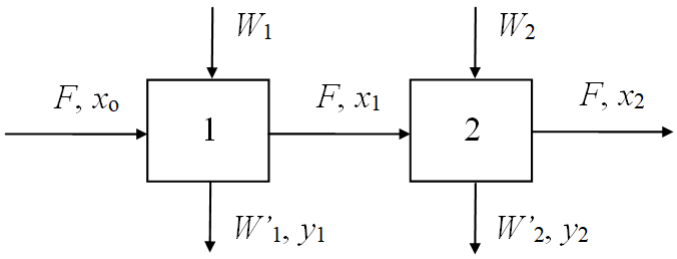


Figura 8.1: Unidade de extração por solvente.

em que W_i e W_i' são vazões mássicas de solvente, F é a vazão mássica de água, x_i é a massa de soluto por unidade de massa de água e y_i é a massa de soluto por unidade de massa de solvente. Uma análise econômica gerou a seguinte expressão do lucro do sistema:

lucro: $L = R - C$
receita: $R = P_s(W_1' y_1 + W_2' y_2)$
custo: $C = P_x(W_1 + W_2)$
restrição: $R > C$

em que P_s é o preço do soluto no extrato, P_x é o preço do solvente puro. Uma análise técnica gerou as seguintes relações restritivas:
balanços de massa para o soluto:

$$Fx_o - W_1' y_1 - Fx_1 = 0$$
$$Fx_1 - W_2' y_2 - Fx_2 = 0$$

balanços de massa para o solvente:

$$W_1 - W_1' - sF = 0$$
$$W_2 + sF - W_2' - sF = 0$$

relações de equilíbrio:

$$y_1 = mx_1$$
$$y_2 = mx_2$$

em que s é a solubilidade do solvente em água (massa de solvente / massa de água) e m é a constante de equilíbrio entre as fases. Portanto, dados F, x_o, s, m, P_s e P_x, o problema de extrair o soluto da água da maneira mais lucrativa possível, consiste em maximizar L em função das condições de operação. Este exemplo possui 6 equações (4 equações de balanço de massa e 2 equações de equilíbrio) e 8 variáveis (W_1, W_2, W_1', W_2', y_1, y_2, x_1 e x_2), tendo, portanto, 2 variáveis de decisão, ou dois graus de liberdade. Eliminando as equações de igualdade e tomando x_1 e x_2 (escolhidas por possuírem os menores intervalos de validade) como variáveis de decisão, a região de busca, ilustrada na Figura 8.2 para o problema de otimização fica delimitada pelas seguintes restrições de desigualdade:

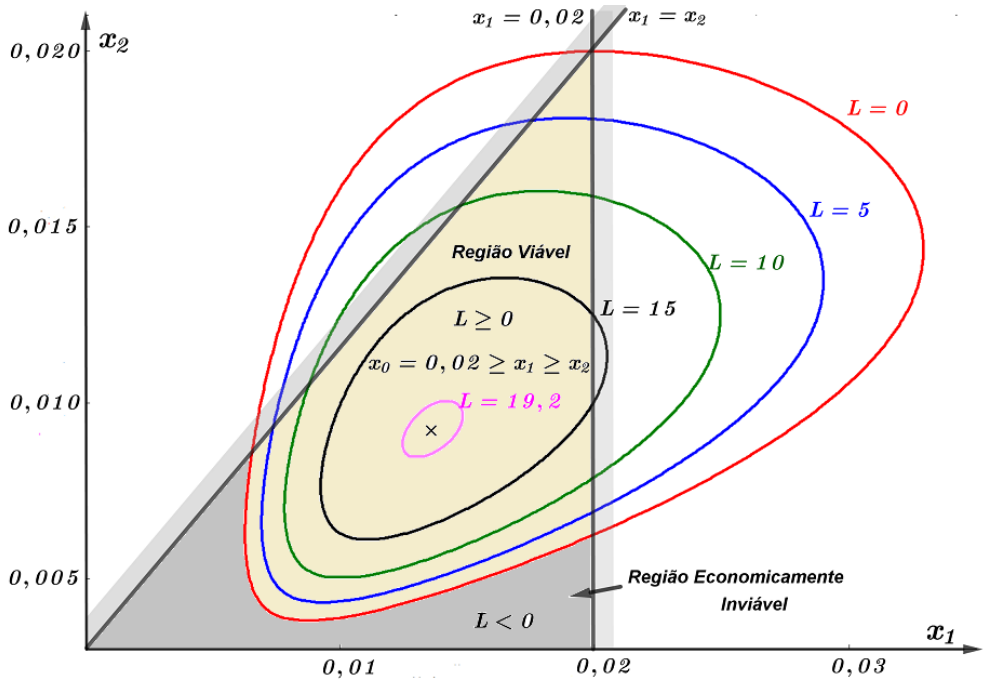


Figura 8.2: Curvas de nível e região viável.

$$x_o \geq x_1 \geq x_2 > 0$$
$$L(x_1, x_2) = a - bx_2 - c/x_1 - dx_1/x_2 > 0$$

em que $a = F[P_s x_o + P_x(2/m - s)]$, $b = FP_s$, $c = FP_x x_o/m$ e $d = FP_x/m$. A Figura 8.2 ilustra a região de busca para $F = 1{,}0 \times 10^4$ kg-água / h, $x_o = 0{,}02$ kg-soluto / kg-água, $s = 7{,}0 \times 10^{-4}$ kg-solvente / kg-água, $m = 4{,}0$ kg-água / kg-solvente, $P_s = 0{,}4$ R$ / kg-soluto e $P_x = 0{,}01$ R$ / kg-solvente.

A formulação do problema de otimização para este exemplo é a seguinte:

$$\max_{x_1,x_2} L(x_1,x_2)$$

sujeito a: $L(x_1,x_2) > 0$
$x_o \geq x_1$
$x_1 \geq x_2$
$x_2 > 0$

■

8.1 Condições de Otimalidade

Seja uma função $S : X \subseteq \Re^n \to \Re$. Diz-se que x^* é um mínimo global (ou absoluto) de S se $S(x^*) \leq S(x) \forall x \in X$, e que x^* é um mínimo local (ou relativo) de S se existe $\varepsilon > 0$, tal que $S(x^*) \leq S(x) \forall x$ tal que $||x - x^*|| < \varepsilon$, ilustrados na Figura 8.3. Se as desigualdades forem estritas, isto é, $S(x^*) < S(x)$ tem-se mínimos globais e locais estritos.

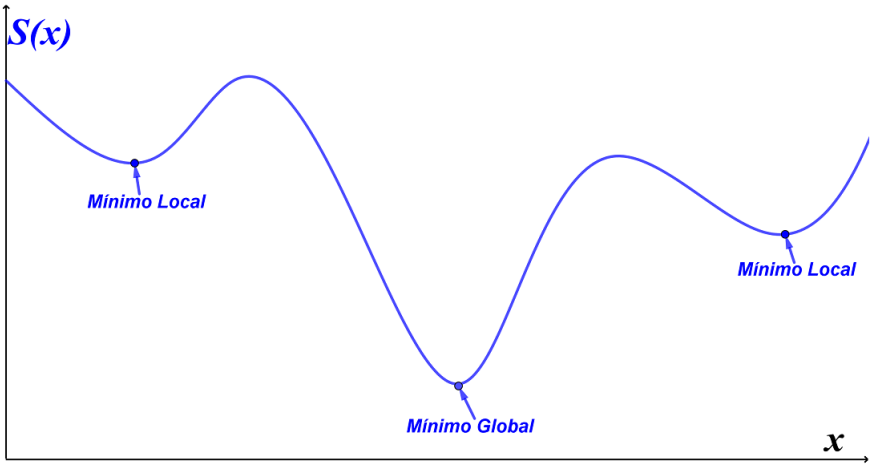


Figura 8.3: Mínimos locais e global.

Um subconjunto K de um espaço vetorial X é dito convexo se $\forall x_1, x_2 \in K$ e $0 \leq \alpha \leq 1$:

$$\alpha x_1 + (1-\alpha)x_2 \in K$$

e apresenta as seguintes propriedades:

a) $\beta K = \{x \in K / x = \beta y, y \in K\}$ é convexo para $\forall \beta \in \Re$;

b) $K + L$ e $K \cap L$ são convexos para $\forall$ subconjunto convexo L de X.

Seja K um convexo não vazio do $\Re^n$. A função $S : K \to \Re$ é dita convexa se $\forall x_1, x_2 \in K$ e $0 \leq \alpha \leq 1$:

$$S[\alpha x_1 + (1-\alpha)x_2] \leq \alpha S(x_1) + (1-\alpha)S(x_2)$$

A função $S(x)$ é estritamente convexa se a desigualdade for estrita. Uma função $T(x)$ é côncava se a função $S(x) = -T(x)$ for convexa.

Sejam $S(x)$, $S_i(x)$, $i = 1, 2, ..., n$, funções convexas em um convexo não vazio K do $\Re^n$, então as seguintes propriedades são apresentadas:

a) $S(x)$ é contínua em qualquer ponto do interior de K;

b) as seções $K_\varepsilon = \{x \in K / S(x) \leq \varepsilon\}$ são conjuntos convexos;

c) $S(x) = \sum_{i=1}^{n} S_i(x)$ é uma função convexa. Se no mínimo uma $S_i(x)$ é estritamente convexa, então $S(x)$ é estritamente convexa;
d) $\beta S(x)$ é convexa para $\beta > 0 \in \Re$.
e) se todas $S_i(x) < \infty \forall x \in K$, então $S(x) = max\{S_1(x), S_2(x), ..., S_n(x)\}$ é convexa.

Se $S(x)$ é convexa então um mínimo local é também global e se a função é estritamente convexa o mínimo global é único.

Para ilustrar, define-se $y = \alpha x_1 + (1-\alpha)x_2$ e $w = \alpha S(x_1) + (1-\alpha)S(x_2)$, representados na Figura 8.4.

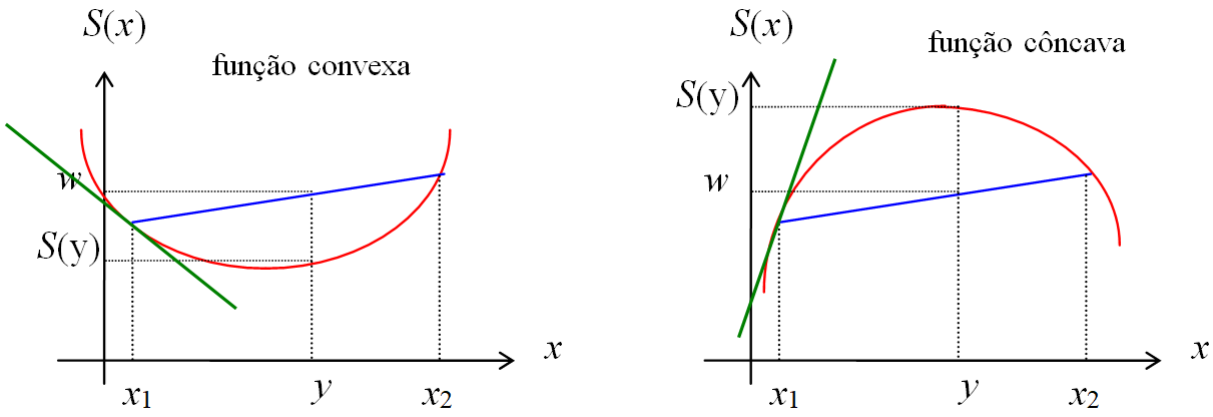


Figura 8.4: Condições de convexidade de ordens zero e um.

Seja X um conjunto aberto não vazio do $\Re^n$ e $S(x)$ uma função diferenciável em $x^o \in X$. Se $S(x)$ é convexa em x^o, então

$$S(x) - S(x^o) \geq \nabla^T S(x^o)(x - x^o).$$

Para uma função $S(x)$ diferenciável em X,

$$S(x) \text{ é convexa} \Leftrightarrow S(x_2) - S(x_1) \geq \nabla^T S(x_1)(x_2 - x_1) \; \forall x_1, x_2 \in X.$$

Seja X um conjunto aberto não vazio do $\Re^n$ e $S(x)$ uma função duas vezes diferenciável em $x^o \in X$. Se $S(x)$ é convexa em x^o, então a matriz das derivadas segunda de $S(x)$, chamada de *matriz Hessiana* em homenagem ao matemático Hesse[1], em x^o, $H(x^o)$, é positiva semidefinida (ver Apêndice A.7). Para uma função $S(x)$ duas vezes diferenciável em X,

$$S(x) \text{ é convexa} \Leftrightarrow H(x) \text{ é positiva semidefinida } \forall x \in X.$$

Seja K um conjunto convexo não vazio do $\Re^n$, $x^o \in K$ e d um vetor não nulo tal que $x^o + \alpha d \in K$ para um $\alpha > 0$ suficientemente pequeno. Então, a *derivada direcional* de $S(x)$ no ponto x^o, ao longo da direção d, denotada por $S'(x^o, d)$, é definida pelo seguinte limite:

$$S'(x^o, d) = \lim_{\alpha \to 0^+} \frac{S(x^o + \alpha d) - S(x^o)}{\alpha} \approx \lim_{\alpha \to 0^+} \left\{ \nabla^T S(x^o) d + \alpha d^T H(x^o) d \right\}.$$

Portanto, a derivada direcional no ponto x^o é dada por $S'(x^o, d) = \nabla^T S(x^o) d$.

[1] Ludwig Otto Hesse (1811-1874).

8.1.1 Otimização sem restrição

No caso da otimização sem restrições, na qual se deseja encontrar os pontos extremos da função objetivo:

$$\min_{x\in\Re^n} S(x) \text{ ou } \max_{x\in\Re^n} S(x)$$

tem-se as seguintes condições de otimalidade.

- Condição necessária de primeira ordem:
  Para que x^* seja um mínimo (máximo) local da função $S(x)$, diferenciável em x^*, é necessário que:
  $$\nabla S(x^*) = 0.$$
- Condição necessária de segunda ordem:
  Para que x^* seja um mínimo (máximo) local da função $S(x)$, duas vezes diferenciável em x^*, é necessário que:
  $$\nabla S(x^*) = 0 \text{ e que}$$
  $$H(x^*) \text{ seja positiva (negativa) semidefinida}$$
  Observa-se que essas condições são apenas necessárias porque os termos de primeira e segunda ordem podem estar nulos, deixando ainda dúvida sobre a natureza de x^*.
- Condição suficiente:
  Seja $S(x)$ duas vezes diferenciável em x^* tal que:
  $$\nabla S(x^*) = 0 \text{ e}$$
  $$H(x^*) \text{ seja positiva (negativa) definida,}$$
  então x^* é um mínimo (máximo) local estrito de S.

Pode-se analisar a condição da matriz Hessiana, $H(x^*)$, pelas seguintes formas:

1) Pela sua contribuição no termo de segunda ordem da expansão em série de Taylor em torno do ponto ótimo.
   $$S(x) - S(x^*) = \frac{1}{2}(x-x^*)^T H(x^*)(x-x^*) + \cdots$$
2) Pelos sinais dos valores característicos de $H(x^*)$.
   Decompondo a matriz Hessiana em seus valores e vetores característicos:
   $$H(x^*) = V\Lambda V^{-1}$$
   em que V é a matriz dos vetores característicos (nas colunas) e Λ é a matriz dos valores característicos (na diagonal). Definindo $z = V^{-1}(x-x^*)$ e lembrando que sendo a matriz Hessiana simétrica então $V^{-1} = V^T$ (matriz ortogonal) e $z^T = (x-x^*)^T V$. Dessa forma a expansão em série de Taylor pode ser escrita como:
   $$S(x) - S(x^*) = \frac{1}{2} z^T \Lambda z + \cdots = \frac{1}{2}\sum_{i=1}^{n} \lambda_i z_i^2 + \cdots$$
3) Pelos sinais dos determinantes das primeiras menores principais de $H(x^*)$ (critério de Sylvester).
   A *menor* M_{ij} de uma matriz H é definida como a matriz obtida pela remoção da i-ésima linha e da j-ésima coluna de H. Uma *menor principal* de ordem k é uma matriz obtida pela remoção de quaisquer $n-k$ colunas e suas linhas correspondentes de uma matriz de ordem n. A *primeira menor principal* de ordem k de uma matriz H, denotada por $M_k(H)$, é obtida pela

remoção das últimas $n-k$ colunas e linhas da matriz H. Observa-se que os determinantes das primeiras menores principais de ordem $1,2,...,n$ da matriz Λ são, respectivamente: $\lambda_1, \lambda_1\lambda_2, ..., \lambda_1\lambda_2\lambda_3\cdots\lambda_n$.

Na tabela a seguir apresentam-se as relações entre a matriz Hessiana e as três formas de analisar a sua condição.

$H(x^*)$	Taylor	λ	$\Delta_k = det(M_k)$
positiva semidefinida	$x^T Hx \geq 0$	≥ 0	≥ 0
positiva definida	$x^T Hx > 0$	> 0	> 0
negativa semidefinida	$x^T Hx \leq 0$	≤ 0	$(-1)^k\Delta_k \geq 0$
negativa definida	$x^T Hx < 0$	< 0	$(-1)^k\Delta_k > 0$
não definida	$x^T Hx \lessgtr 0$	$\lessgtr 0$	$\neq$ dos acima

Dessa forma, pode-se afirmar que x^* é:

- ponto de mínimo local se $H(x^*)$ for positiva definida, $S(x) > S(x^*)$
- ponto de máximo local se $H(x^*)$ for negativa definida, $S(x) < S(x^*)$
- ponto de sela se $H(x^*)$ for não definida, ora $S(x) > S(x^*)$ e ora $S(x) < S(x^*)$

Nas situações em que $H(x^*)$ é semidefinida deve-se ainda investigar os termos de ordem superior da expansão em série de Taylor.

Se $S(x)$ é convexa então as condições de otimalidade simplificam-se, porque as condições de segunda ordem são equivalentes à convexidade local da função.

■ **Exemplo 8.2 — Função monovariável.** $S(x) = (x^2-1)^3$

$\nabla S(x) = 6x(x^2-1)^2 \rightarrow \nabla S(x) = 0 \begin{cases} x_1 = 0 \\ x_2 = 1 \\ x_3 = -1 \end{cases}$, satisfazem a condição necessária de primeira ordem;

$H(x) = (x^2-1)(30x^2-6) \rightarrow H(x_1) = 6; H(x_2) = 0; H(x_3) = 0$ satisfazem a condição necessária de segunda ordem; contudo somente x_1 satisfaz a condição suficiente. Nesse caso (univariável) x_2 e x_3 são pontos de inflexão, como pode ser visto na Figura 8.5, ou avaliando o valor da derivada terceira de $S(x)$ nesses pontos:

$$\nabla^3 S(x) = 24x(5x^2-3) \rightarrow \nabla^3 S(x_2) = 48 \neq 0 \text{ e } \nabla^3 S(x_3) = -48 \neq 0.$$

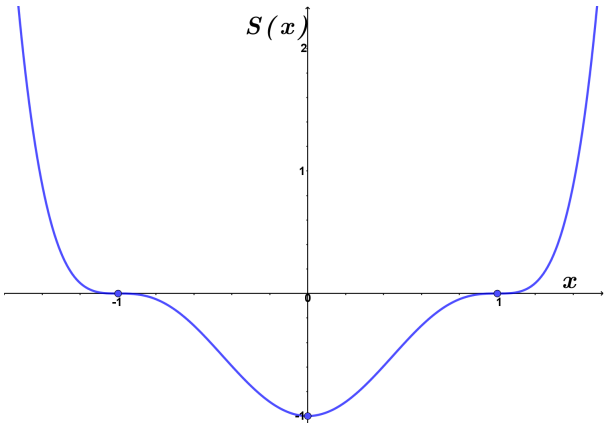


Figura 8.5: Função monovariável.

■

■ **Exemplo 8.3 — Função multivariável.** $S(x_1,x_2) = x_1^2 + x_1 x_2^2$

$$\nabla S(x) = \begin{bmatrix} 2x_1 + x_2^2 \\ 2x_1 x_2 \end{bmatrix}$$

então $x_1^* = x_2^* = 0$, e:

$$H(x) = \begin{bmatrix} 2 & 2x_2 \\ 2x_2 & 2x_2 \end{bmatrix} \rightarrow H(x*) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

isto é, uma matriz positiva semidefinida. A Figura 8.6 ilustra a função $S(x)$, na qual se observa que $x^* = 0$ não é um ponto de mínimo. Fazendo a mesma análise com a mudança de variável $y = x_2^2$, verifica-se que a origem é um ponto sela.

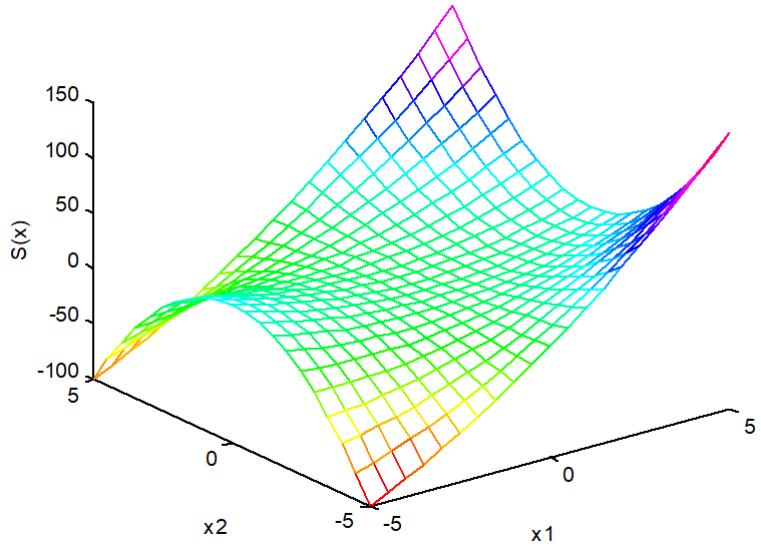


Figura 8.6: Função multivariável.

■

8.1.2 Otimização com restrições

Seja o problema de otimização sujeito a restrições de igualdade, $h(x)$, e desigualdade, $g(x)$:

$$\begin{aligned} &\min_{x \in X} S(x) \\ &\text{sujeito a:} \quad h_j(x) = 0, j = 1,2,...,m \\ &\qquad\qquad\ \ g_j(x) \leq 0, j = 1,2,...,p \\ &\qquad\qquad\ \ x \in X \subseteq \Re^n \end{aligned}$$

em que $S(x)$, $g(x)$ e $h(x) \in C^2$. O conjunto de todos os pontos viáveis é definido por:

$$K = \{x \in X \subseteq \Re^n / h(x) = 0, g(x) \leq 0\}.$$

Uma restrição de desigualdade $g_j(x)$ é chamada de *ativa* em um ponto viável $\bar{x}$ se $g_j(\bar{x}) = 0$, caso contrário ela é uma *restrição inativa*. As restrições ativas restringem a região de viabilidade, enquanto que as inativas não impõem restrição alguma na vizinhança do ponto $\bar{x}$, definida pela hiperesfera de raio ε em torno desse ponto, denotada por $B_\varepsilon(\bar{x})$.

Um vetor d é chamado de *vetor de direção viável* a partir do ponto $\bar{x}$ se existe uma hiperesfera de raio ε tal que:

$$(\bar{x} + \alpha d) \in \{B_\varepsilon(\bar{x}) \cap K\} \text{ para todo } 0 \leq \alpha \leq \frac{\varepsilon}{\|d\|}.$$

O conjunto de vetores de direções viáveis a partir de $\bar{x}$ é chamado de *cone de direções viáveis* de K no ponto $\bar{x}$. A Figura 8.7 ilustra essas definições.

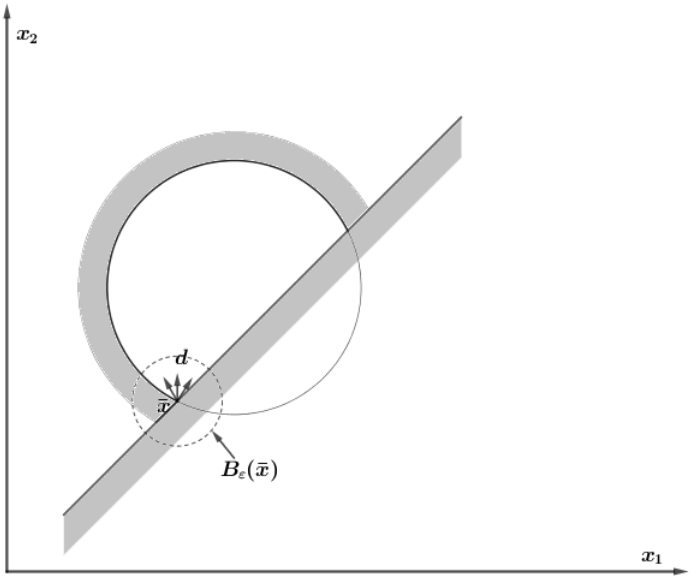


Figura 8.7: Cone de direções viáveis.

Se $d \neq 0$, então $\bar{x}$ deve satisfazer as seguintes condições:
$d^T \nabla h(\bar{x}) = 0$
$d^T \nabla g(\bar{x}) \leq 0$ para as restrições ativas, pois $g(x) \approx g(\bar{x}) + \alpha \nabla^T g(\bar{x}) d \leq 0$
e se $d^T \nabla S(\bar{x}) < 0$, então d é uma direção viável e promissora, isto é,

$$S(\bar{x} + \alpha d) < S(\bar{x}) \text{ para todo } 0 \leq \alpha \leq \frac{\varepsilon}{\|d\|}, \text{ pois } S(x) - S(\bar{x}) \approx \alpha \nabla^T S(\bar{x}) d < 0.$$

Se $\bar{x} = x^*$ é um ponto de mínimo local do problema, então para um α suficientemente pequeno, tem-se:

$$S(x^*) \leq S(x^* + \alpha d).$$

A ideia chave para desenvolver as condições necessárias e suficientes para um problema de otimização com restrições é transformá-lo em um problema de otimização sem restrições e aplicar as condições para este caso. Uma forma de fazer esta transformação é através da introdução de uma função auxiliar, chamada de função de Lagrange, $L(x, \eta, \mu)$, definida como:

$$L(x, \eta, \mu) = S(x) + \eta^T h(x) + \mu^T g(x)\,, \mu \geq 0$$

em que η e μ são os multiplicadores de Lagrange associados às restrições de igualdade e desigualdade, respectivamente (μ são também conhecidos como multiplicadores de Kuhn-Tucker). Desse modo, o problema transformado torna-se:

$$\max_{\eta, \mu \geq 0} \min_{x} L(x, \eta, \mu)$$

em que os multiplicadores η associados com as restrições de igualdade, $h(x) = 0$, assumem sinais positivos quando $H(x) \geq 0$ e negativos quando $h(x) \leq 0$. No ponto ótimo tem-se:

$$L(x^*, \eta^*, \mu^*) = S(x^*).$$

Cada multiplicador de Lagrange indica o quão sensível é a função objetivo em relação à restrição associada. Por exemplo, se as restrições de igualdade são perturbadas por um vetor b, isto é, $h_b(x) = b$, então:

$$\nabla_b S(x^*) = -\eta^*.$$

Note que neste caso a função de Lagrange tem a forma:

$$L(x,\eta) = S(x) + \eta^T [h_b(x) - b]$$

e sua sensibilidade em relação ao parâmetro b é dada por:

$$\nabla_b L = \nabla_b^T x[\nabla_x S(x) + \nabla_x^T h_b(x)\eta] + \nabla_b^T \eta [h_b(x) - b] - \eta.$$

Como os termos entre colchetes da expressão acima devem ser nulos no ponto ótimo (condição necessária de primeira ordem), então:

$$\nabla_b L(x^*,\eta^*) = -\eta^* \text{ e}$$

$$\nabla_b S(x^*) = -\eta^*$$

$$\text{pois } \nabla_b L(x^*,\eta^*) = \nabla_b S(x^*) + \nabla_b^T \eta [h_b(x^*) - b] + \nabla_b^T h_b(x^*)\eta^* - \eta^*,$$

$$h_b(x^*) = b \text{ e } \nabla_b h_b(x^*) = I.$$

Portanto, o valor de $S(x)$ aumenta ou diminui a partir de $S(x^*)$ com um aumento ou diminuição em b, dependendo do sinal de η^*. Por isso, os multiplicadores de Lagrange são também conhecidos como "*shadow prices*" ou "custos marginais" das restrições, porque a mudança no valor ótimo da função objetivo por unidade de acréscimo no lado direito da restrição de igualdade é dado por η^*.

■ **Exemplo 8.4 — Perturbação.** Seja o seguinte problema de otimização com restrição
$\min_{x\in\Re^2} S(x) = (x_1-5)^2 + (x_2-5)^2$
sujeito a: $h(x) = x_1 - x_2 = 0$
Introduzindo uma perturbação na restrição de igualdade do tipo: $x_1 - x_2 = b$, a função de Lagrange toma a forma:

$$L(x,\eta) = (x_1-5)^2 + (x_2-5)^2 + \eta(x_1 - x_2 - b)$$

cujo gradiente nulo com relação à x_1, x_2 e η leva ao seguinte sistema de equações:

$$\begin{aligned} 2(x_1-5) + \eta &= 0 \\ 2(x_2-5) - \eta &= 0 \\ x_1 - x_2 - b &= 0 \end{aligned}$$

resultando na solução ótima: $x_1^* = 5 + b/2$, $x_2^* = 5 - b/2$ e $\eta^* = -b$.
Deste modo $S(x^*) = b^2/2$ e $\nabla_b S(x^*) = b = -\eta^*$. ■

Para entender a origem da função de Lagrange, o ótimo do exemplo acima deve satisfazer as seguintes condições:

$$\begin{aligned} dS &= \tfrac{\partial S}{\partial x_1}\delta x_1 + \tfrac{\partial S}{\partial x_1}\delta x_2 = 0 \\ dh &= \tfrac{\partial h}{\partial x_1}\delta x_1 + \tfrac{\partial h}{\partial x_2}\delta x_2 = 0 \end{aligned}.$$

Se $S(x)$ fosse uma função sem restrição, então as suas duas derivadas parciais seriam nulas no ponto ótimo e $dS(x^*)$ seria nulo para quaisquer valores das variações δx_1 e δx_2. Entretanto, como as variáveis x_1 e x_2 estão restritas (δx_1 e δx_2 não são independentes), as duas derivadas parciais de $S(x)$ não podem ser arbitrariamente igualadas a zero. Contudo, $S(x)$ deve ser um extremo no ponto ótimo e portanto $dS(x^*) = 0$. A segunda condição, $dh(x^*) = 0$, existe porque $h(x) = 0$. Para se

obter uma solução (δx_1 e δx_2) não trivial do sistema de equações acima, a matriz dos coeficientes do sistema:

$$\begin{bmatrix} \frac{\partial S}{\partial x_1} & \frac{\partial S}{\partial x_2} \\ \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} \end{bmatrix}$$

deve ter determinante nulo, ou seja, as linhas da matriz são linearmente dependentes:

$$\frac{\partial S}{\partial x_1} + \eta \frac{\partial h}{\partial x_1} = 0 \text{ e } \frac{\partial S}{\partial x_2} + \eta \frac{\partial h}{\partial x_2} = 0.$$

Então, definindo uma função auxiliar: $L(x,\eta) = S(x) + \eta^T h(x)$
as condições acima são satisfeitas se: $\nabla_x L(x,\eta) = 0$. Para que a restrição de igualdade, $h(x) = 0$, seja também satisfeita é necessário que $\nabla L(x,\eta) = 0$. Portanto, no ponto ótimo é necessário que $\nabla L(x^*,\eta^*) = 0$.

A existência dos multiplicadores de Lagrange depende da forma das restrições, e estará garantida se e somente se os gradientes das restrições de desigualdade ativas, $\nabla g_j(x^*)$, e das restrições de igualdade, $\nabla h(x^*)$, forem linearmente independentes. Por exemplo, no caso de um problema somente com restrições de igualdade, a condição necessária de primeira ordem para $L(x,\eta)$ fica:

$$\nabla_x S(x) + [\nabla_x h(x)]^T \eta = 0$$

cuja solução para η existirá somente se a matriz $\nabla_x h(x)$ possuir posto completo, m, isto é, estar composta por m vetores linearmente independentes.

Dessa forma, para problemas de otimização com restrições, tem-se as seguintes condições de otimalidade.

- Condição necessária de primeira ordem de Karush[2]-Kuhn[3]-Tucker[4] (KKT):
  Para que x^* seja um ótimo local do problema com restrições, com $S(x)$, $g(x)$ e $h(x)$ diferenciáveis em x^*, é necessário que:
  os gradientes das restrições de desigualdade ativas, $\nabla g_j(x^*)$, e das restrições de igualdade, $\nabla h(x^*)$, sejam linearmente independentes (*qualificação de segunda ordem* das restrições), e que as seguintes condições sejam satisfeitas:

  $\nabla_x L(x^*,\eta^*,\mu^*) = \nabla S(x^*) + (\eta^*)^T \nabla h(x^*) + (\mu^*)^T \nabla g(x^*) = 0$
  $h(x^*) = 0$
  $g(x^*) \leq 0$
  $\mu_j^* g_j(x^*) = 0, j = 1,2,...,p$ (*condições de complementaridade*)
  $\mu^* \geq 0$

  A condição do gradiente nulo, $\nabla_x L(x^*,\eta^*,\mu^*) = 0$, implica em:

  $$(\eta^*)^T \nabla h(x^*) + (\mu^*)^T \nabla g(x^*) = -\nabla S(x^*)$$

  que interpretada graficamente, Figura 8.8, mostra que o vetor $\nabla S(x^*)$ pertence ao cone das direções viáveis, formado pelo negativo dos gradientes das restrições de igualdade e desigualdade ativas (uma vez que $\mu_j^* = 0$ para as restrições inativas).
  Supondo que $\nabla S(x^*)$ caia fora do cone das direções viáveis, então haveria uma direção d tal que $d^T \nabla S(x^*) < 0$, $d^T \nabla g(x^*) \leq 0$ e $d^T \nabla h(x^*) = 0$, isto é, existiria um ponto melhor que x^*, como ilustra a Figura 8.9.

[2] William Karush (1917–1997).
[3] Harold William Kuhn (1925–2014).
[4] Albert William Tucker (1905—1995).

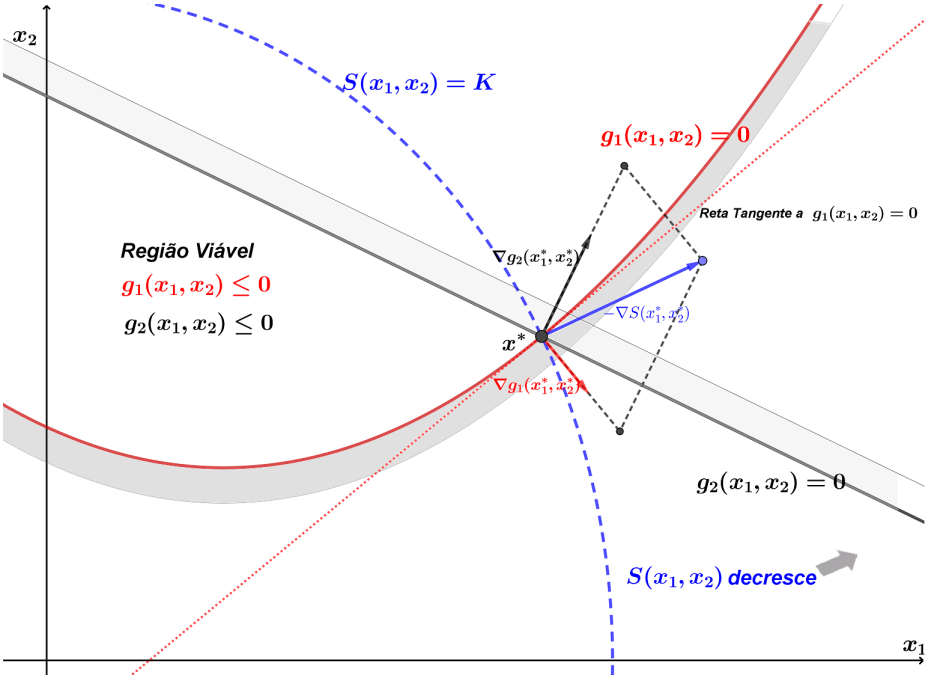


Figura 8.8: Condição de KKT de primeira ordem.

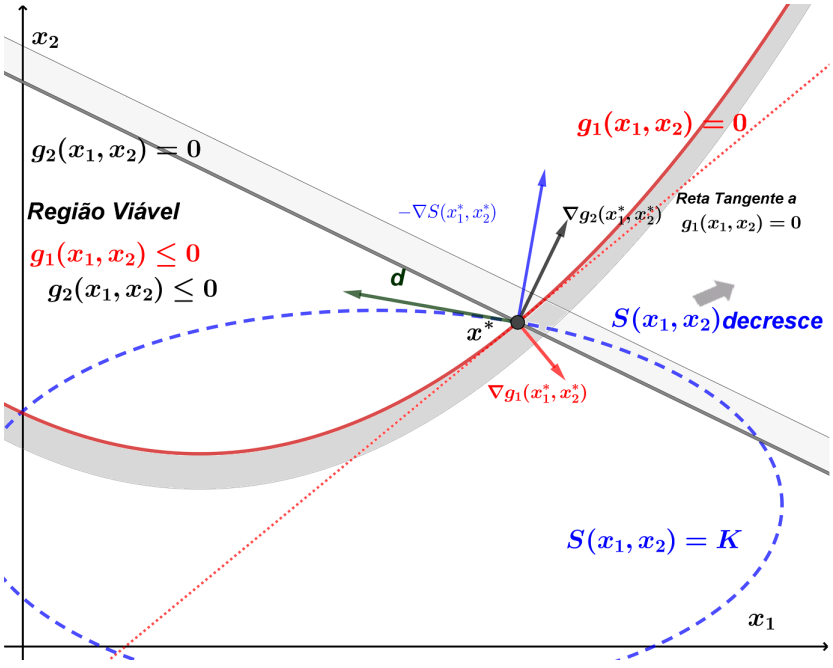


Figura 8.9: Violação da condição de KKT de primeira ordem.

- Condição necessária de segunda ordem de KKT:
  Para que x^* seja um mínimo local do problema com restrições, com $S(x)$, $g(x)$ e $h(x)$ duas vezes diferenciáveis em x^*, é necessário que a condição de primeira ordem de KKT seja satisfeita e que a matriz Hessiana da função de Lagrange, $H_x(x^*, \eta^*, \mu^*)$, seja positiva semidefinida (o subscrito x indica a matriz das derivadas segundas de L em relação a x) para todo vetor não nulo d tal que:

$d^T \nabla h_i(x^*) = 0,\ i = 1,2,...,m$
$d^T \nabla g_j(x^*) = 0,$ para as $g_j(x^*)$ ativas

ou seja, $d^T H_x(x^*, \eta^*, \mu^*) d \geq 0$.

- Condição suficiente de KKT:
  Para que x^* seja um mínimo local do problema com restrições, com $S(x)$, $g(x)$ e $h(x)$ duas vezes diferenciáveis em x^*, é suficiente que a condição de primeira ordem de KKT seja satisfeita e, que a matriz Hessiana da função de Lagrange, $H_x(x^*, \eta^*, \mu^*)$, seja positiva definida para todo vetor não nulo d tal que:

  $d^T \nabla h_i(x^*) = 0, i = 1,2,...,m$
  $d^T \nabla g_j(x^*) = 0$ para as $g_j(x^*)$ ativas $\{g_j(x^*) = 0$ e $\mu_j^* > 0\}$
  $d^T \nabla g_j(x^*) \leq 0$ para as $g_j(x^*)$ inativas $\{g_j(x^*) < 0$ e $\mu_j^* = 0\}$

  ou seja, $d^T H_x(x^*, \eta^*, \mu^*) d > 0$.

A positividade da matriz Hessiana com restrição, isto é:

$$d^T H_x(x^*, \eta^*, \mu^*) d > 0 \forall d \in \{d / d^T \nabla h_i(x^*) = 0, d^T \nabla g_j(x^*) = 0, d \neq 0\}$$

é garantida se todas as raízes do polinômio característico

$$p(\lambda) = \begin{vmatrix} \lambda I - H_x & M \\ M^T & 0 \end{vmatrix} = 0$$

forem positivas, em que M é a matriz formada pelos gradientes de $h(x^*)$ e $g(x^*)$ ativas, isto é, a matriz tal que $d^T M = 0$, com $m + p^a < n$ e com posto completo (p^a é o número de restrições g ativas). O mesmo critério se aplica para semipositividade, negatividade e seminegatividade, com os respectivos sinais das raízes.

■ **Exemplo 8.5 — KKT 1.** Verificar as condições necessárias e suficientes para o seguinte problema (Edgar, Himmelblau e Lasdon, 2001):

$\min_{x \in \Re^2} S(x) = (x_1 - 1)^2 + x_2^2$
sujeito a: $g_1(x) = x_1 - x_2^2/4 \leq 0$ ■

■ **Exemplo 8.6 — KKT 2.** Verificar as condições necessárias e suficientes para o problema com a mesma função objetivo do exemplo anterior, mas usando a seguinte restrição:

$g_2(x) = x_1 - x_2^2 \leq 0$

■

8.2 Métodos Diretos

Neste texto abordaremos somente alguns métodos de otimização sem restrição, cujo problema a ser resolvido é:

$$\min_{x \in \Re^n} S(x) \text{ ou } \max_{x \in \Re^n} S(x)$$

como $\max_{x \in \Re^n} S(x)$ é equivalente a $\min_{x \in \Re^n} -S(x)$, os métodos descritos a seguir são para problemas de minimização. Os métodos existentes para a resolução desse problema podem ser agrupados em duas categorias:

1) métodos que não usam derivadas (métodos de busca, métodos diretos);

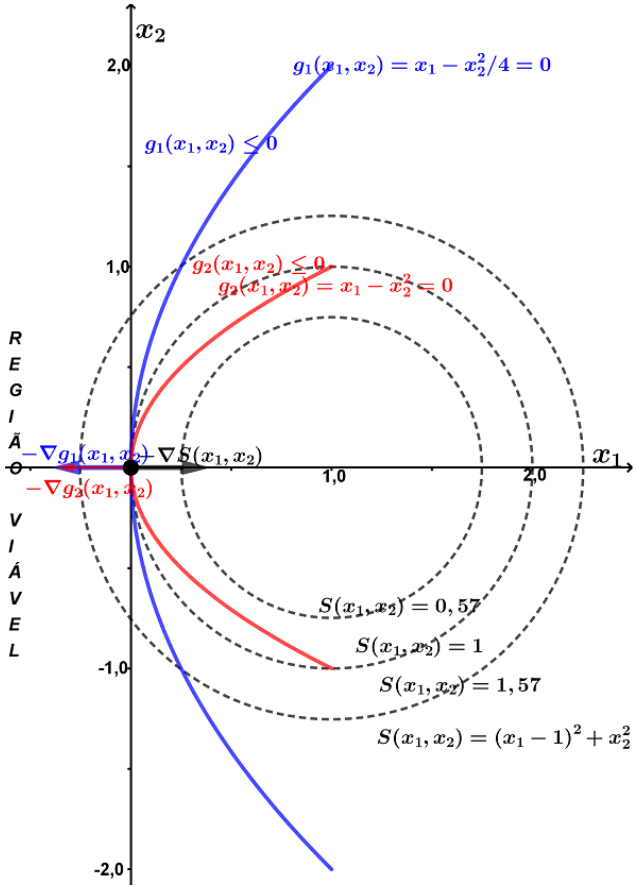


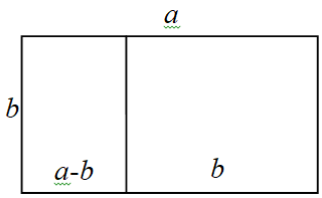
Figura 8.10: Exemplos 8.5 e 8.6 da condição de KKT de primeira ordem.

2) métodos que usam derivadas (métodos analíticos, métodos da métrica variável, métodos indiretos).

Como regra geral, na resolução de problemas sem restrição, os métodos que usam derivadas convergem mais rapidamente que os métodos de busca. Por outro lado, os métodos de busca não requerem regularidade e continuidade da função objetivo e, principalmente o cálculo de derivadas primeira ou segunda de $S(x)$.

8.2.1 Método da Seção Áurea

É um método de busca monovariável, no qual a cada iteração o intervalo de busca é reduzido por um fator α, chamado de razão áurea, obtido pela relação geométrica da figura abaixo (retângulo áureo):



$$\frac{a-b}{b} = \frac{b}{a}$$

$$\left(\frac{b}{a}\right)^2 + \frac{b}{a} - 1 = 0$$

$$\alpha = \frac{b}{a} = \frac{-1+\sqrt{5}}{2} = 0,618$$

Outras escolhas do fator α levariam a métodos similares, como por exemplo o método da bisseção para $\alpha = 0,5$. Contudo, a vantagem da razão áurea está na redução do número de cálculos da função objetivo, em função da propriedade deste método de conservar o retângulo áureo a cada

iteração. Outro método com características similares a seção áurea é a busca de **Fibonacci**[5].

Algoritmo 8.1 — Seção Áurea.

1) Determinar o intervalo de busca $[L^o, U^o]$ que contém o ponto de mínimo
2) Fazer $k \leftarrow 0$, $\Delta^o \leftarrow U^o - L^o$,
$x_L^o \leftarrow L^o + \alpha\Delta^o$, $S_L^o \leftarrow S(x_L^o)$,
$x_U^o \leftarrow U^o - \alpha\Delta^o$, $S_U^o \leftarrow S(x_U^o)$,
3) **Se** $S_U^k > S_L^k$, **então** $L^{k+1} \leftarrow x_U^k$, $x_U^{k+1} \leftarrow x_L^k$, $S_U^{k+1} \leftarrow S_L^k$
senão $U^{k+1} \leftarrow x_L^k$, $x_L^{k+1} \leftarrow x_U^k$, $S_L^{k+1} \leftarrow S_U^k$
4) Fazer $k \leftarrow k+1$ e $\Delta^k \leftarrow U^k - L^k$,
Se $S_U^{k-1} > S_L^{k-1}$, **então** $x_L^k \leftarrow L^k + \alpha\Delta^k$, $S_L^k \leftarrow S(x_L^k)$
senão $x_U^k \leftarrow U^k - \alpha\Delta^k$, $S_U^k \leftarrow S(x_U^k)$
5) **Se** $\Delta^k > \varepsilon$ (tolerância), **então** (ir para 3)
senão FIM.

A Figura 8.11 ilustra duas iterações sucessivas do método seção áurea.

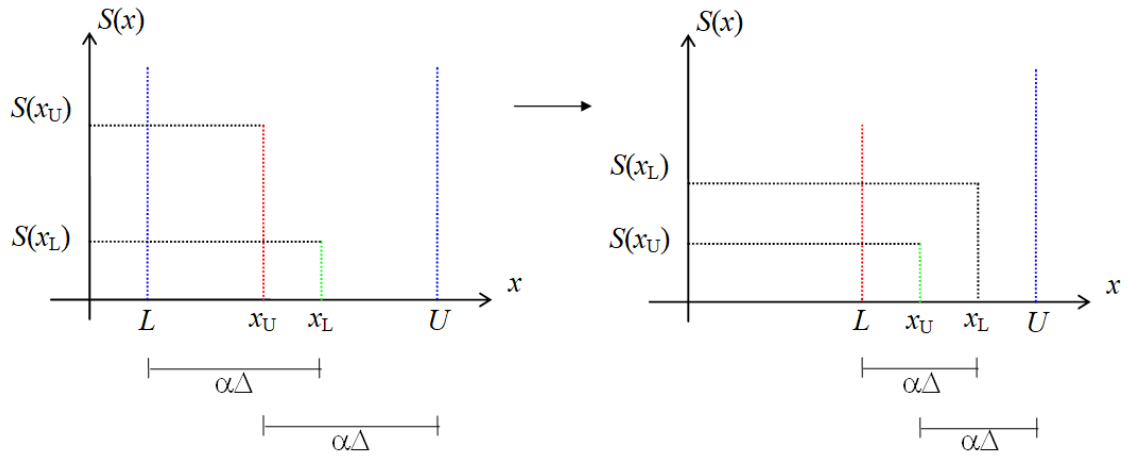


Figura 8.11: Método da seção áurea.

8.2.2 Método das Aproximações Polinomiais Sucessivas

É uma classe de métodos de busca monovariável que aproxima a função $S(x)$ por uma interpolação polinomial, $P_n(x)$:

$$P_n(x) = \sum_{j=1}^{n} \ell_j(x) S(x_j)$$

em que $\ell_j(x) = \prod_{k=1 \neq j}^{n} \frac{(x - x_k)}{(x_j - x_k)}$ são os interpoladores de Lagrange.

Quando a derivada de $S(x)$ está disponível, então ela também é usada na aproximação polinomial:

$$P_{2n-1}(x) = \sum_{j=1}^{n} [h_1(x) S(x_j) + h_2(x) \nabla S(x_j)]$$

[5] Leonardo Pisano Fibonacci (1170–1250).

em que $h_1(x) = \ell_j^2(x)\,[1-2(x-x_j)\nabla\ell_j(x_j)]$ e $h_2(x) = \ell_j^2(x)(x-x_j)$ são os interpoladores de Hermite.

Interpolação quadrática ou método de Coggins (ou DSC-Powell)
O método de Coggins (1964) ou de Davies-Swann-Campey-Powell[6] (Swann, 1972) aproxima a função $S(x)$ por uma interpolação quadrática, $P_2(x)$:

$$P_2(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}S(x_1) + \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}S(x_2) + \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}S(x_3)$$

calculando o mínimo desta função quadrática, isto é, $\frac{dP_2}{dx} = 0$, tem-se:

$$x^{\#} = \frac{1}{2}\frac{(x_2^2-x_3^2)S(x_1) + (x_3^2-x_1^2)S(x_2) + (x_1^2-x_2^2)S(x_3)}{(x_2-x_3)S(x_1) + (x_3-x_1)S(x_2) + (x_1-x_2)S(x_3)}$$

Algoritmo 8.2 — Coggins.

1) Determinar o intervalo de busca $[x_1, x_3]$
2) Calcular $S_1 \leftarrow S(x_1)$, $S_3 \leftarrow S(x_3)$, $x_2 \leftarrow (x_1+x_3)/2$ e $S_2 \leftarrow S(x_2)$
3) Calcular $x^{\#} \leftarrow \frac{1}{2}\frac{(x_2^2-x_3^2)S_1+(x_3^2-x_1^2)S_2+(x_1^2-x_2^2)S_3}{(x_2-x_3)S_1+(x_3-x_1)S_2+(x_1-x_2)S_3}$ e $S^{\#} \leftarrow S(x^{\#})$
4) **Se** $|x^{\#}-x_i| < \varepsilon$ para algum $i = 1,2,3$, **então** FIM.
5) **Se** $(x_3-x^{\#})(x^{\#}-x_2) > 0$, **então** $k \leftarrow 1$
   senão $k \leftarrow 3$
6) **Se** $S^{\#} < S_2$, **então** $x_k \leftarrow x_2$, $S_k \leftarrow S_2$ e $k \leftarrow 2$
7) $x_{4-k} \leftarrow x^{\#}$, $S_{4-k} \leftarrow S^{\#}$ e (ir para 3).

A Figura 8.12 ilustra a determinação de $x^{\#}$ e $S^{\#}$ decorrente da primeira aproximação parabólica de $S(x)$. Johnson e Townsend (1978) avaliaram o desempenho e condições de falha do método de Coggins.

8.2.3 Método de Hooke & Jeeves

É um método de busca multivariável proposto por Hooke e Jeeves (1961) dividido em duas fases, ilustrado na Figura 8.13:

Fase de exploração:	estimar a direção provável do extremo, a partir de um ponto inicial (ponto base).
Fase de progressão:	progredir na direção provável do extremo enquanto o valor da função objetivo estiver diminuindo.

Algoritmo 8.3 — Hooke & Jeeves.

Partida:

1) Determinar a região de busca $[L_i, U_i]$ $(i = 1,2,\ldots,n)$
2) Selecionar o ponto base inicial x_{io} $(i = 1,2,\ldots,n)$
3) Calcular o valor $S_o \leftarrow S(x_o)$ da função objetivo em x_o
4) Selecionar os incrementos iniciais δ_i e as respectivas tolerâncias ε_i $(i = 1,2,\ldots,n)$
5) Tomar a primeira direção de busca $(k \leftarrow 1)$

Fase de Exploração:

6) Calcular $x_{ko}^{-} \leftarrow x_{ko} - \delta_k$ (sentido negativo)
7) **Se** x_{ko}^{-} estiver fora da região de busca, **então** insucesso em k^{-} (ir para 10)
8) Calcular o valor de $S_o^{-} \leftarrow S(x_o^{-})$
9) **Se** $S_o^{-} > S_o$, **então** insucesso em k^{-}
   senão sucesso em k^{-} e fazer $x_{ko} \leftarrow x_{ko}^{-}$ e $S_o \leftarrow S_o^{-}$ (ir para 14)

[6] Michael James David Powell (1936–2015).

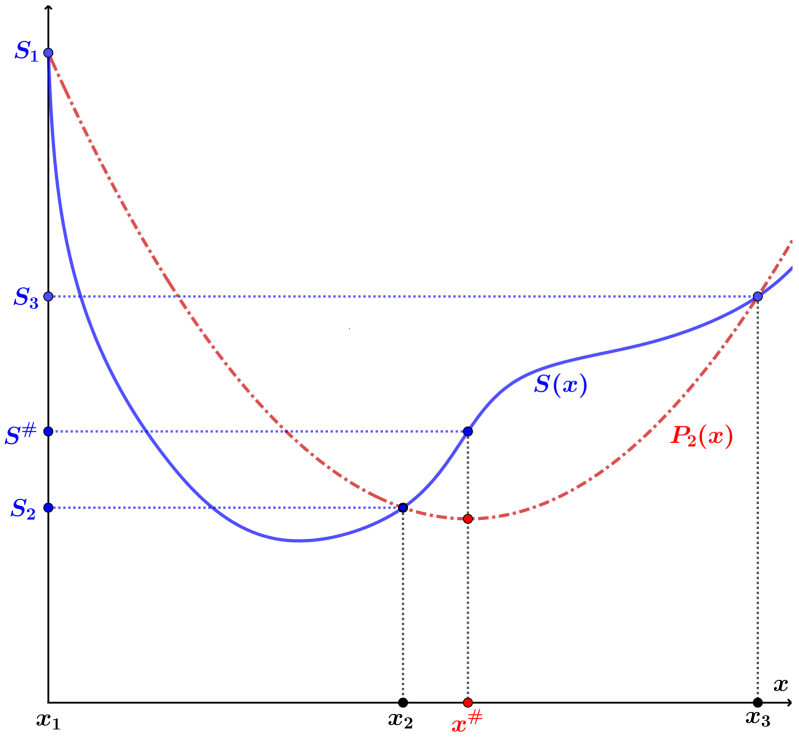


Figura 8.12: Método de Coggins.

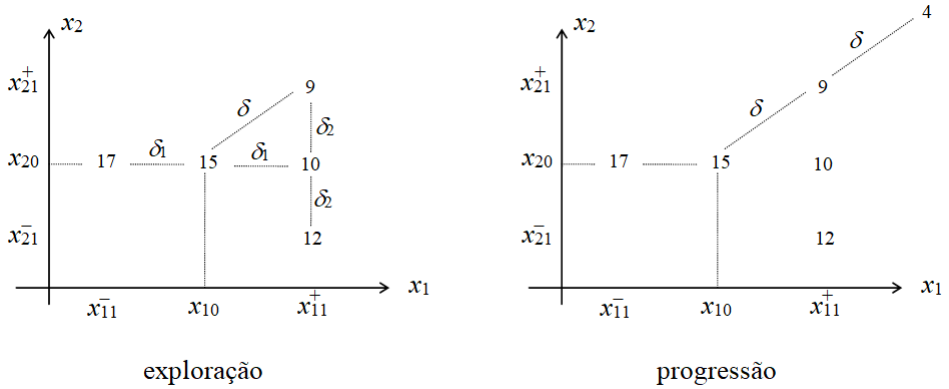


Figura 8.13: Método de Hooke & Jeeves.

10) Calcular $x_{ko}^+ \leftarrow x_{ko} + \delta_k$ (sentido positivo)
11) **Se** x_{ko}^+ estiver fora da região de busca, **então** insucesso em k^+ (ir para 14)
12) Calcular o valor de $S_o^+ \leftarrow S(x_o^+)$
13) **Se** $S_o^+ > S_o$, **então** insucesso em k^+
senão sucesso em k^+ e fazer $x_{ko} \leftarrow x_{ko}^+$ e $S_o \leftarrow S_o^+$
14) **Se** já foram exploradas todas as direções ($k = n$), **então** (ir para 15)
senão tomar a direção seguinte
$k \leftarrow k+1$ e (ir para 6)
15) **Se** houve sucesso em alguma direção **então** (ir para 17)
16) **Se** $\delta_i \leq \varepsilon_i \; \forall i$, **então** FIM.
senão $\delta_i \leftarrow \delta_i/2 \; \forall i$ tal que $\delta_i > \varepsilon_i$ (ir para 5)

Fase de Progressão:

17) Tomar $x_{i1} \leftarrow x_{io} \pm \delta_i$ (e, opcionalmente, $\delta_i \leftarrow 2\delta_i$) para todas as direções que houve sucesso
18) **Se** x_1 estiver fora da região de busca, **então** insucesso na progressão (ir para 16)
19) Calcular o valor de $S_1 \leftarrow S(x_1)$
20) **Se** $S_1 > S_o$, **então** insucesso na progressão (ir para 5)
senão sucesso na progressão e fazer $x_o \leftarrow x_1$ e $S_o \leftarrow S_1$ (ir para 17)

8.2.4 Método de Busca de Limites

A maioria dos métodos de busca necessita da definição do intervalo de busca, que contém o ponto extremo, para garantirem a determinação do ótimo. Segue abaixo um algoritmo de busca para a determinação de um intervalo $[x_o, x_1]$ que contém um ponto de mínimo.

Algoritmo 8.4 — Busca de Limites.

1) Escolher um ponto inicial, x_o, e um passo inicial, α, calcular $S_o \leftarrow S(x_o), k \leftarrow 0$
2) Calcular $x_1 \leftarrow x_o + \alpha$ e $S_1 \leftarrow S(x_1)$
3) **Se** $S_o \leq S_1$, **então** $\alpha \leftarrow -\alpha, k \leftarrow k+1$ e (ir para 6)
4) $S_o \leftarrow S_1, \alpha \leftarrow 2\alpha, x_1 \leftarrow x_o + \alpha$ e $S_1 \leftarrow S(x_1)$
5) **Se** $S_o > S_1$, **então** (ir para 4)
senão $k \leftarrow 1$ (ir para 7)
6) **Se** $k < 2$, **então** (ir para 2)
7) $x_o \leftarrow x_o + \alpha(k-1), I \leftarrow [x_o, x_1]$, FIM.

8.2.5 Método dos Poliedros Flexíveis

É um método de busca multivariável no qual o pior vértice de um poliedro com $n+1$ vértices é substituído por um novo vértice colinear com o vértice antigo e o centroide (Nelder e Mead, 1965), como ilustra a Figure 8.14.

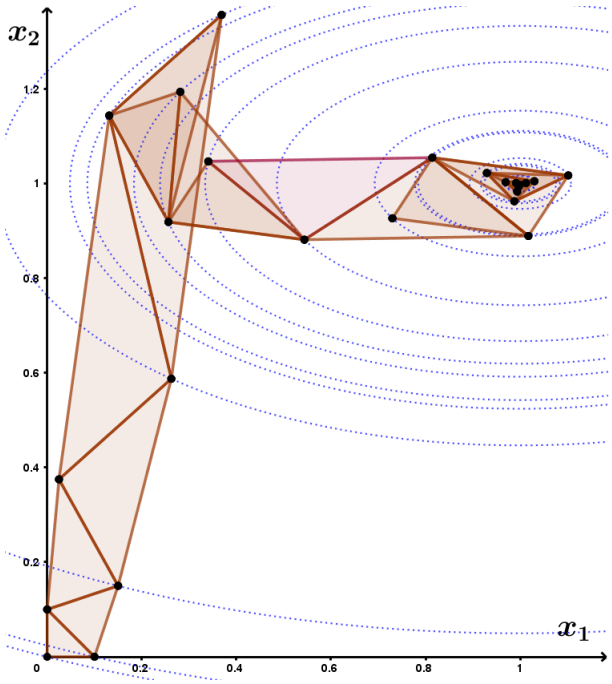


Figura 8.14: Método dos poliedros flexíveis.

As coordenadas do centroide do poliedro são dadas por:

$$x_{0,j} = \frac{1}{n}\left[\sum_{i=1}^{n+1} x_{i,j} - x_{h,j}\right], \; j = 1,2,...,n$$

em que $x_{h,j}$ é o pior vértice.

O algoritmo envolve quatro operações de busca, que para o caso da minimização da função objetivo têm as seguintes formas:

1) Reflexão: $\begin{cases} x_R^k \leftarrow x_0^k + \alpha(x_0^k - x_h^k),\ \alpha > 0 \\ \text{em que } S(x_h^k) \leftarrow max\{S(x_1^k),...,S(x_{n+1}^k)\} \end{cases}$

2) Expansão:
$$\begin{cases} \textbf{Se } S(x_R^k) \leq S(x_\ell^k) = min\{S(x_1^k),...,S(x_{n+1}^k)\}, \\ \textbf{então} \\ \qquad x_E^k \leftarrow x_0^k + \gamma(x_R^k - x_0^k),\ \gamma > 1 \\ \qquad \textbf{Se } S(x_E^k) \leq S(x_R^k), \textbf{então } x_h^{k+1} \leftarrow x_E^k \\ \qquad\qquad\qquad\qquad \textbf{senão } x_h^{k+1} \leftarrow x_R^k \\ \qquad k \leftarrow k+1 \text{ (ir para 1)} \end{cases}$$

em que x_ℓ^k é o melhor vértice.

3) Contração:
$$\begin{cases} \textbf{Se } S(x_R^k) \leq S(x_i^k)\ \forall i \neq h, \\ \textbf{então} \\ \qquad x_C^k \leftarrow x_0^k + \beta(x_h^k - x_0^k),\ 0 < \beta < 1 \\ \qquad x_h^{k+1} \leftarrow x_C^k \\ \qquad k \leftarrow k+1 \text{ (ir para 1)} \end{cases}$$

4) Redução:
$$\begin{cases} \textbf{Se } S(x_R^k) > S(x_h^k), \\ \textbf{então} \\ \qquad x_i^{k+1} \leftarrow x_\ell^k + \frac{1}{2}(x_i^k - x_\ell^k),\ i = 1,2,...,n+1 \\ \qquad k \leftarrow k+1 \text{ (ir para 1)} \end{cases}$$

O critério usado por Nelder[7] e Mead[8] para terminar a busca é o seguinte:

$$\left\{\frac{1}{n+1}\sum_{i=1}^{n+1}\left[S(x_i^k) - S(x_0^k)\right]^2\right\}^{\frac{1}{2}} \leq \varepsilon.$$

8.2.6 Métodos Não Determinísticos

São métodos nos quais caminhos aleatórios são construídos na sequência dos passos em busca do ótimo. Existe uma grande variedade de métodos que se enquadram nesta categoria, tais como os algoritmos genéticos, busca aleatória adaptativa, *simulated annealing*, PSO, etc. O *Particle Swarm Optimization* (PSO), proposto por Kennedy e Eberhart (1995), é um algoritmo que tem como fundamento o comportamento de organismos sociais tais como uma revoada de pássaros ou um cardume de peixes, em que cada indivíduo da população (partícula) modifica sua posição com o tempo (geração). A velocidade e posição de uma partícula são modificadas de acordo com a experiência do indivíduo (valor da função objetivo) e a dos demais componentes da população, valendo-se de sua melhor posição e a melhor posição do conjunto, dadas por:

$$v_i^{k+1} = wv_i^k + c_1\alpha_i^{k+1}(p_i^k - x_i^k) + c_2\beta_i^{k+1}(g^k - x_i^k)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1}$$

em que x_i^k é a posição da partícula i na iteração k, v_i^k sua velocidade, p_i^k sua melhor posição até a iteração k, g^k é a melhor posição dentre todas as partículas até a iteração k, α_i^k e β_i^k são números

[7] John Ashworth Nelder (1924–2010).
[8] Roger Mead (1938–2015).

aleatórios entre 0 e 1, w (fator de inércia) e c_1 e c_2 são parâmetros de sintonia do método. Para valores elevados de w, tem-se uma melhor busca global e para valores pequenos têm-se uma melhor busca local. $c_2 > c_1$ favorece o refinamento da solução já obtida, ou seja, a busca local, ao passo que $c_2 < c_1$ favorece a busca global, já que cada partícula segue o seu próprio caminho.

Algoritmo 8.5 — PSO.

1) Inicialização: construção de x_i^0 e v_i^0 aleatoriamente, $k = 0$.
2) Avalia-se a aptidão de cada partícula e atualiza-se p_i^k e g^k.
3) Calcula-se a nova velocidade v_i^{k+1} e posição $x_i^{k+1}, k \leftarrow k+1$
4) **Se** $k <$ *número máximo de iterações* **então** (ir para 2)

No passo (4) pode-se também adotar outros critérios de parada, como por exemplo, um determinado número de iterações sem haver mudança no valor de g^k.

8.3 Métodos Indiretos

Esses métodos têm como equação básica para o processo iterativo:

$$x^{k+1} = x^k - \alpha_k W(x^k) \nabla S(x^k)$$

em que α_k é o tamanho do passo, $d^k = -W(x^k)\nabla S(x^k)$ é o vetor direção e $W(x^k)$ é a matriz direção (inversa da matriz Hessiana ou uma aproximação dessa).

Em qualquer método de otimização, uma boa direção de busca deve reduzir (para o caso da minimização) o valor da função objetivo, isto é, $S(x^{k+1}) < S(x^k)$. Tal direção, d^k, satisfaz o seguinte critério em cada ponto:

$$\nabla^T S(x^k) d^k < 0$$

ou em outras palavras, o ângulo (θ) formado entre os vetores $\nabla S(x^k)$ e d^k deve ser sempre maior que 90^o, conforme ilustrado na Figura 8.15, ou seja:

$$\nabla^T S(x^k) d^k = |\nabla^T S(x^k)||d^k| cos(\theta) < 0 \Leftrightarrow \theta > 90^o$$

Como a otimização sem restrições é equivalente a encontrar a solução do sistema de equações não lineares $F(x) = \nabla S(x) = 0$ (daí vem a origem do nome dos métodos indiretos), pode-se utilizar todos os métodos disponíveis para a solução de $F(x) = 0$. Por exemplo, na utilização do método de Newton-Raphson, a matriz Jacobiana é a própria matriz Hessiana.

8.3.1 Método do Gradiente

Utilizam somente a primeira derivada da função objetivo, caso em que $W(x^k) = I$, Figura 8.16, sendo uma adaptação do método das substituições sucessivas (Seção 4.3) para otimização:

$$x^{k+1} = x^k - \alpha_k \nabla S(x^k)$$

Quando α_k é escolhido de modo a minimizar:

$$g_k(\alpha) = S[x^k - \alpha \nabla S(x^k)],\ \alpha > 0$$

tem-se o método da descida mais íngreme (*steepest descent*), ilustrado na Figura 8.17, cujo algoritmo básico pode ser escrito da seguinte forma.

Algoritmo 8.6 — *Steepest Descent*.

1. Escolher um ponto inicial x^o, $k \leftarrow 0$

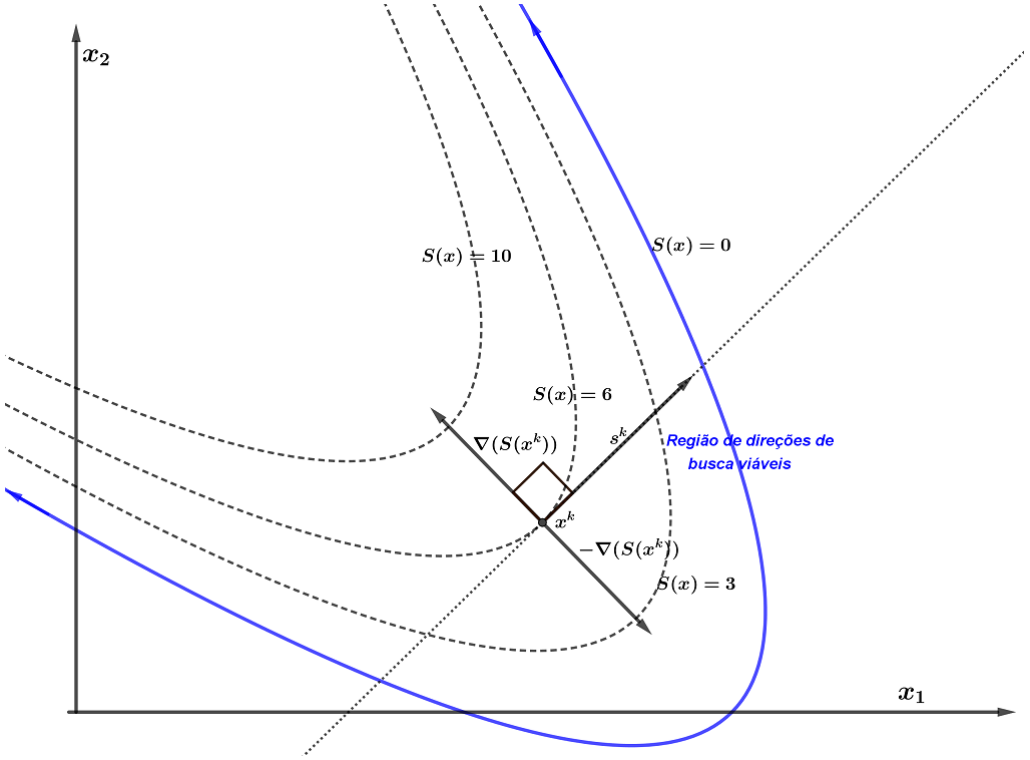


Figura 8.15: Direções de busca.

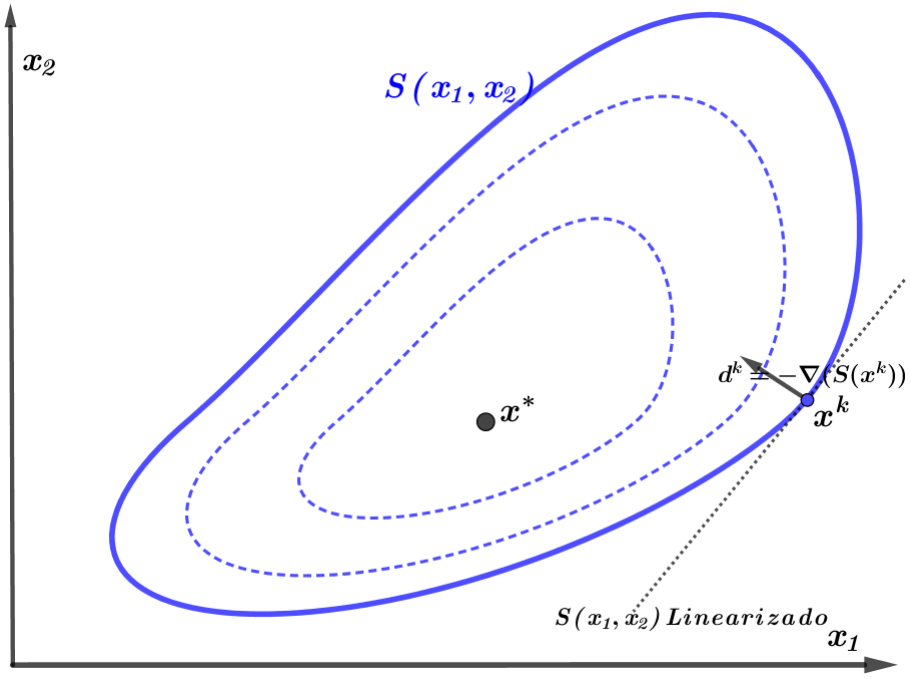


Figura 8.16: Método Gradiente.

2. Calcular $d^k \leftarrow -\nabla S(x^k)$
3. Encontrar α_k tal que $S(x^k + \alpha_k d^k) = \min_{\alpha>0} g_k(\alpha) = S(x^k + \alpha d^k)$
4. Calcular $x^{k+1} \leftarrow x^k + \alpha_k d^k$

5. **Se** o critério de convergência não foi satisfeito, **então** $k \leftarrow k+1$ (ir para 2)
6. FIM.

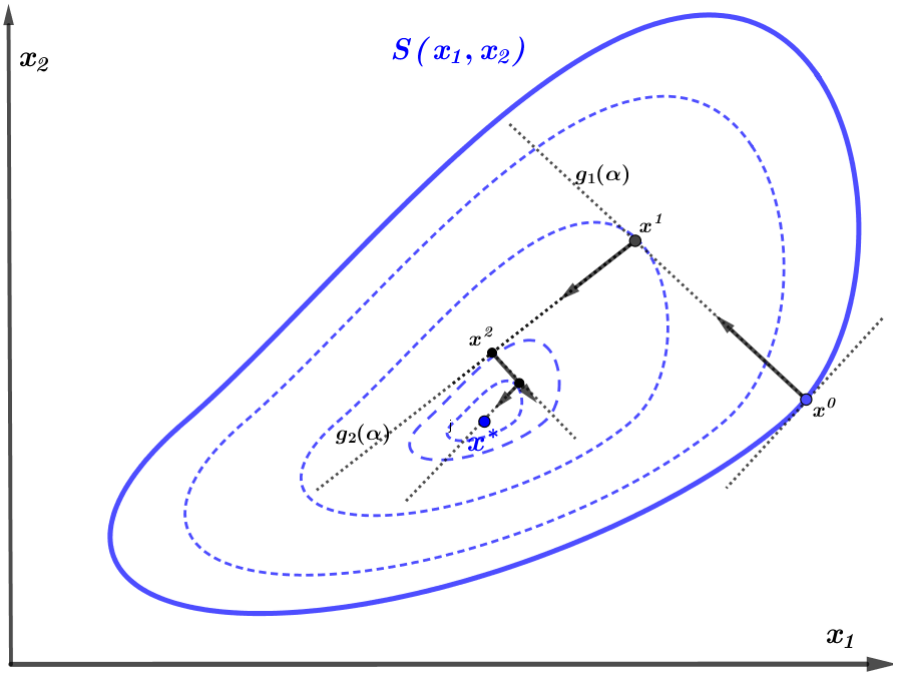


Figura 8.17: Método da descida mais íngreme.

A minimização de $g_k(\alpha)$, conhecida como *função de mérito*, também chamada de busca em linha (*linesearch*), pode ser realizada com o uso de qualquer método de minimização univariável. Para ilustrar esta função, na Figura 8.18 é mostrada a função $g_1(\alpha)$ do problema ilustrado na Figura 8.17.

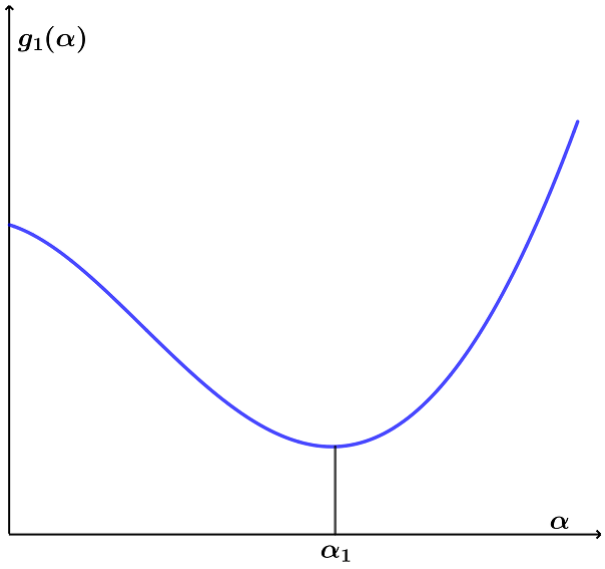


Figura 8.18: Método da descida mais íngreme.

Aproximando $S(x)$ por uma função quadrática:

$$S(x^{k+1}) \approx S(x^k) + \nabla^T S(x^k)(x^{k+1} - x^k) + \frac{1}{2}(x^{k+1} - x^k)^T H(x^k)(x^{k+1} - x^k)$$

ou de forma similar:

$$g_k(\alpha) = S(x^k + \alpha d^k) \approx S(x^k) + \alpha \nabla^T S(x^k) d^k + \frac{1}{2}\alpha^2 (d^k)^T H(x^k) d^k$$

que minimizando em relação a α, ou seja, $\frac{dg_k}{d\alpha} = 0$, resulta:

$$\alpha^* = \alpha_k = -\frac{\nabla^T S(x^k) d^k}{(d^k)^T H(x^k) d^k} = \frac{(d^k)^T d^k}{(d^k)^T H(x^k) d^k}$$

Contudo, a equação acima não é utilizada para o cálculo de α no método do gradiente, pois exigiria o cálculo da segunda derivada da função objetivo. Nesse caso, se utiliza, em geral, métodos de busca para a sua seleção.

8.3.2 Método de Newton

Faz uso da segunda derivada da função objetivo, caso em que $W(x^k) = H(x^k)^{-1}$:

$$x^{k+1} = x^k - \alpha_k [H(x^k)]^{-1} \nabla S(x^k)$$

que é resultado da minimização da aproximação de $S(x)$ por uma função quadrática, ilustrada na Figura 8.19:

$$S(x) \approx S(x^k) + \nabla^T S(x^k)\,\Delta x^k + \frac{1}{2}(\Delta x^k)^T H(x^k)\,\Delta x^k,$$

em que $\Delta x^k = x - x^k$, na direção Δx^k, isto é, $\frac{\partial S}{\partial \Delta x_i^k} = 0$:

$$\Delta x^k = -[H(x^k)]^{-1} \nabla S(x^k).$$

Nesse caso α_k ou é um parâmetro de relaxação do processo iterativo $0 < \alpha_k \leq 1$, ou é um fator de correção da inversa da matriz Hessiana, caso esta não seja atualizada em todas as iterações. A positividade da matriz Hessiana deve estar sempre garantida para evitar a migração para um ponto sela ou ponto de máximo. E, para assegurar a convergência do método de Newton, a correção Δx^k deve ser tal que $S(x^{k+1}) < S(x^k)$.

Uma maneira de assegurar a positividade da matriz Hessiana é através da modificação de Levenberg[9]-Marquardt[10], que adiciona um fator ajustável na diagonal da matriz Hessiana, ou em sua inversa:

$$\tilde{H}(x^k) = H(x^k) + \beta_k I \ , \ \beta_k > -min\{\lambda_i\}$$
$$W(x^k) = [H(x^k)]^{-1} + \gamma_k I \ , \ \gamma_k > -min\{1/\lambda_i\}$$

em que λ_i são os valores característicos de $H(x^k)$.

Em particular, quando o método de Newton é utilizado para a solução de problemas de mínimos quadrados, ele é comumente referenciado na literatura como método de Gauss-Newton. Sendo que uma das aplicações é na resolução de sistemas de equações não lineares, $F(x) = 0$, transformados em problemas de mínimos quadrados ao procurar minimizar o quadrado dos resíduos, isto é,

$$S(x) = F^T(x)F(x) = \sum_{i=1}^{m} f_i^2(x)$$

[9] Kenneth Levenberg (1919–1973).
[10] Donald W. Marquardt (1929–1997).

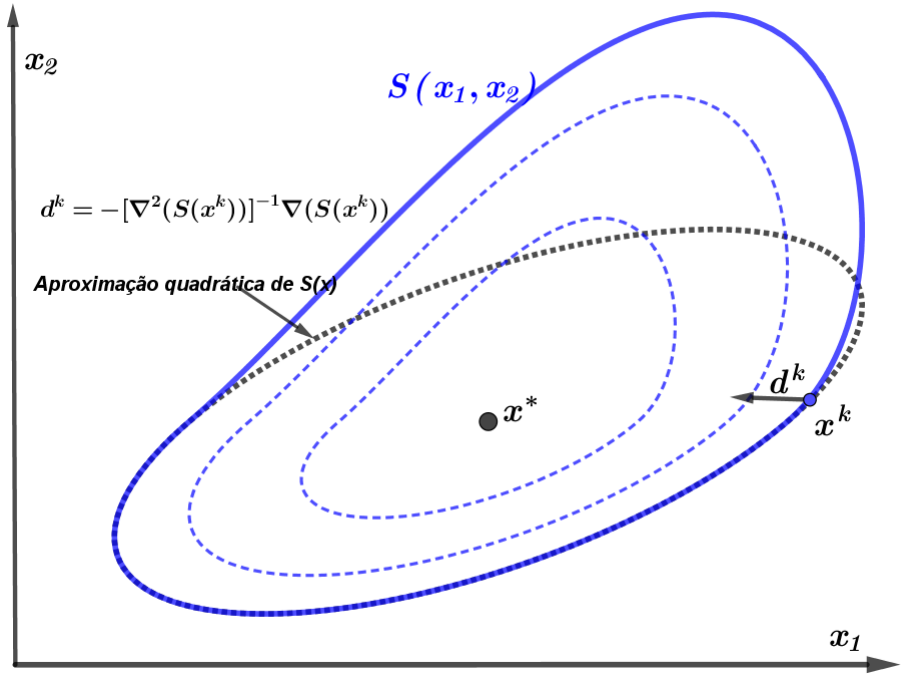


Figura 8.19: Método da Newton.

neste caso $\nabla S(x^k) = 2J^T(x^k)F(x^k)$ e $H(x^k) = 2J^T(x^k)J(x^k) + 2Q(x^k)$, em que $J(x) = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j}$ é a matriz Jacobiana do sistema, $Q(x) = \sum_{i=1}^{m} f_i(x)H_i(x)$ e $H_i(x)$ é a matriz Hessiana da função $f_i(x)$. Quando $f_i(x^k) \to 0$ para $x^k \to x^*$, então $Q(x^k)$ tende a zero, e as direções de busca do método de Gauss-Newton para o problema de mínimos quadrados:

$$\min_{d\in\Re^n} \left\|J(x^k)d - F(x^k)\right\|^2$$

são equivalentes às direções do método de Newton, ou seja:

$$d^k = -[J^T(x^k)J(x^k)]^{-1}J^T(x^k)F(x^k) \approx -[H(x^k)]^{-1}\nabla S(x^k).$$

8.3.3 Método do Gradiente Conjugado

Utiliza somente a primeira derivada da função objetivo, gerando uma sequência de direções que são combinações lineares do gradiente:

$$d^{k+1} = \varepsilon_{k+1}d^k - \nabla S(x^{k+1})$$

em que a nova direção é conjugada com a direção anterior com respeito a Hessiana:

$$(d^{k+1})^T H(x^k)d^k = 0$$

e $x^{k+1} = x^k + \alpha_k d^k$, em que α_k é obtido de forma similar ao método da maior descida. Para calcular ε_{k+1}, faz-se a aproximação quadrática de $S(x)$, da qual obtém-se:

$$\nabla S(x) \approx \nabla S(x^k) + H(x^k)(x - x^k)$$

e, portanto: $\nabla S(x^{k+1}) - \nabla S(x^k) = H(x^k)(x^{k+1} - x^k) = H(x^k)\alpha_k d^k$, que multiplicado por d^{k+1} à esquerda, resulta:

$$(d^{k+1})^T[\nabla S(x^{k+1}) - \nabla S(x^k)] = \alpha_k(d^{k+1})^T H(x^k)d^k = 0$$

e, substituindo a equação para d^{k+1} na expressão acima, tem-se:

$$[\varepsilon_{k+1}d^k - \nabla S(x^{k+1})]^T[\nabla S(x^{k+1}) - \nabla S(x^k)] = 0,$$

mas devido à ortogonalidade entre a direção de busca e o gradiente da função objetivo no mínimo desta direção $(d^k)^T\nabla S(x^{k+1}) = 0$ e para a aproximação quadrática $\nabla^T S(x^{k+1})\nabla S(x^k) = 0$, resulta em:

$$\varepsilon_{k+1} = -\frac{\nabla^T S(x^{k+1})\,\nabla S(x^{k+1})}{(d^k)^T\,\nabla S(x^k)} = \frac{\nabla^T S(x^{k+1})\,\nabla S(x^{k+1})}{\nabla^T S(x^k)\,\nabla S(x^k)}.$$

A última igualdade resulta do fato de $d^k = \varepsilon_k d^{k-1} - \nabla S(x^k)$, que multiplicado por $\nabla S(x^k)$ à direita:

$$(d^k)^T\nabla S(x^k) = \varepsilon_k(d^{k-1})^T\nabla S(x^k) - \nabla^T S(x^k)\nabla S(x^k) = -\nabla^T S(x^k)\nabla S(x^k),$$

pois $(d^{k-1})^T\nabla S(x^k) = 0$, pela mesma razão acima.

Algoritmo 8.7 — Gradiente Conjugado.

1) Escolher um ponto inicial x^o
2) Calcular $d^o \leftarrow -\nabla S(x^o), k \leftarrow 0$
3) Encontrar α_k tal que $S(x^k + \alpha_k d^k) = \min_{\alpha>0} g_k(\alpha) = S(x^k + \alpha d^k)$
4) Calcular $x^{k+1} \leftarrow x^k + \alpha_k d^k$ e $\nabla S(x^{k+1})$
5) **Se** o critério de convergência foi satisfeito, **então** FIM.
6) Calcular $d^{k+1} \leftarrow -\nabla S(x^{k+1}) + d^k\dfrac{\nabla^T S(x^{k+1})\nabla S(x^{k+1})}{\nabla^T S(x^k)\nabla S(x^k)}, k \leftarrow k+1$
7) **Se** $k = n$, isto é, realizou n direções L.I. **então** fazer $x^o \leftarrow x^k$ e (ir para 2)
   senão (ir para 3).

8.4 Método dos Mínimos Quadrados

Seja uma relação envolvendo m variáveis independentes $x_1, x_2, \ldots, x_m$ com uma variável dependente y:

$$y_{mod}(\mathbf{x}, \mathbf{a}) = \sum_{i=0}^{n} a_i f_i(\mathbf{x}),$$

sendo $a_0, a_1, ..., a_n$ os chamados $n+1$ parâmetros do *modelo*. Nesse caso diz-se que a função é ***linear*** nos parâmetros. Assim: $\dfrac{\partial y_{mod}(\mathbf{x}, \mathbf{a})}{\partial a_k} = f_k(\mathbf{x})$ para $k = 0, 1, ..., n$.

Deseja-se determinar os n parâmetros do modelo que *minimizem* a *função objetivo* que é a *soma dos quadrados dos erros*:

$$S(\mathbf{a}) = \sum_{j=1}^{N_{exp}} [y_{exp,j} - y_{mod}(\mathbf{x}_j, \mathbf{a})]^2 = \sum_{j=1}^{N_{exp}} \left[y_{exp,j} - \sum_{i=0}^{n} a_i f_i(\mathbf{x}_j) \right]^2.$$

Logo: $\dfrac{\partial S(\mathbf{a})}{\partial a_k} = -2\sum_{j=1}^{N_{exp}} f_k(\mathbf{x}_j) \left[y_{exp,j} - \sum_{i=0}^{n} a_i f_i(\mathbf{x}_j) \right] = 0,$

ou seja, $\sum_{i=0}^{n} \left[\sum_{j=1}^{N_{exp}} f_k(\mathbf{x}_j) f_i(\mathbf{x}_j) \right] a_i = \sum_{j=1}^{N_{exp}} f_k(\mathbf{x}_j)\, y_{exp,j}$

Adotando a notação matricial: $\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \in \Re^{n+1}$; $\mathbf{y}_{exp} = \begin{pmatrix} y_{exp,1} \\ y_{exp,2} \\ \vdots \\ y_{exp,N_{exp}} \end{pmatrix} \in \Re^{N_{exp}}$;

$$\mathbf{A} = \begin{pmatrix} f_0(\mathbf{x}_1) & f_1(\mathbf{x}_1) & \cdots & f_n(\mathbf{x}_1) \\ f_0(\mathbf{x}_2) & f_1(\mathbf{x}_2) & \cdots & f_n(\mathbf{x}_2) \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ f_0(\mathbf{x}_{N_{exp}}) & f_1(\mathbf{x}_{N_{exp}}) & \cdots & f_n(\mathbf{x}_{N_{exp}}) \end{pmatrix}.$$

Assim: $\left[\mathbf{A}^T\mathbf{A}\right]_{linha\,k} = \sum_{j=1}^{N_{exp}} f_k(\mathbf{x}_j)\, f_i(\mathbf{x}_j)$ e $\left[\mathbf{A}^T\mathbf{y}_{exp}\right]_{elemento\,k} = \sum_{j=1}^{N_{exp}} f_k(\mathbf{x}_j)\,\mathbf{y}_{exp,j}$

Resultando no sistema linear:

$$\left[\mathbf{A}^T\mathbf{A}\right]\mathbf{a} = \mathbf{A}^T\mathbf{y}_{exp} \Rightarrow \mathbf{a} = \left[\mathbf{A}^T\mathbf{A}\right]^{-1}\left(\mathbf{A}^T\mathbf{y}_{exp}\right).$$

Uma forma de ajuste bastante empregada é o *ajuste polinomial*, nesse caso: $f_i(x) = x^i$ para $i = 0, 1, ..., n$, assim:

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^n \\ 1 & x_2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 1 & x_{N_{exp}} & \cdots & x_{N_{exp}}^n \end{pmatrix}$$

No caso particular de $n = 1$ tem-se o *ajuste linear* quando:

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{N_{exp}} \end{pmatrix} \text{ e } \mathbf{A}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{N_{exp}} \end{pmatrix}$$

Assim: $\mathbf{A}^T\mathbf{A} = \begin{pmatrix} N_{exp} & \sum_{i=1}^{N_{exp}} x_i \\ \sum_{i=1}^{N_{exp}} x_i & \sum_{i=1}^{N_{exp}} x_i^2 \end{pmatrix}$ e $\mathbf{A}^T\mathbf{y}_{exp} = \begin{pmatrix} \sum_{i=1}^{N_{exp}} y_{exp,i} \\ \sum_{i=1}^{N_{exp}} x_i\, y_{exp,i} \end{pmatrix}$ ou seja:

$$\begin{pmatrix} N_{exp} & \sum_{i=1}^{N_{exp}} x_i \\ \sum_{i=1}^{N_{exp}} x_i & \sum_{i=1}^{N_{exp}} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N_{exp}} y_{exp,i} \\ \sum_{i=1}^{N_{exp}} x_i\, y_{exp,i} \end{pmatrix},$$

dividindo ambos os membros por N_{exp} e definindo os *valores médios*:

$$\langle x \rangle = \frac{\sum_{i=1}^{N_{exp}} x_i}{N_{exp}},\ \langle y_{exp} \rangle = \frac{\sum_{i=1}^{N_{exp}} y_{exp,i}}{N_{exp}},\ \langle x^2 \rangle = \frac{\sum_{i=1}^{N_{exp}} x_i^2}{N_{exp}} \text{ e } \langle x\, y_{exp} \rangle = \frac{\sum_{i=1}^{N_{exp}} x_i\, y_i}{N_{exp}}$$

resulta em: $\begin{pmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \langle y_{exp} \rangle \\ \langle x\, y_{exp} \rangle \end{pmatrix}.$

Quando a equação do modelo é ***não linear*** nos parâmetros, pode se proceder da seguinte maneira (método de Gauss-Newton): considerando $\mathbf{a}^{(k)}$ o valor do vetor dos parâmetros na iteração k, lineariza-se a equação do modelo em torno de $\mathbf{a}^{(k)}$, resultando em:

$$y_{mod,linearizado}(\mathbf{x},\mathbf{a}) = y_{mod}(\mathbf{x},\mathbf{a}^{(k)}) + \sum_{i=0}^{n}\left[a_i - a_i^{(k)}\right] f_i^{(k)}(\mathbf{x}) = \hat{y}_{mod}^{(k)}(\mathbf{x}) + \sum_{i=0}^{n} a_i\, f_i^{(k)}(\mathbf{x}),$$

sendo $f_i^{(k)}(\mathbf{x}) = \left.\frac{\partial y_{mod}(\mathbf{x},\mathbf{a})}{\partial a_i}\right|_{\mathbf{a}^{(k)}}$ e $\hat{y}_{mod}^{(k)}(\mathbf{x}) = y_{mod}(\mathbf{x},\mathbf{a}^{(k)}) - \sum_{i=0}^{n} a_i^{(k)}\, f_i^{(k)}(\mathbf{x})$.

O valor de $\mathbf{a}^{(k+1)}$ na próxima iteração $(k+1)$ é então calculado por:

$$\mathbf{a}^{(k+1)} = \left[\mathbf{A}_k^T\,\mathbf{A}_k\right]^{-1}\left(\mathbf{A}_k^T\,\mathbf{y}_{exp}^{(k)}\right)$$

sendo: $\mathbf{y}_{exp}^{(k)} = \begin{pmatrix} y_{exp,1} - \hat{y}_{mod}^{(k)}(\mathbf{x}_1) \\ y_{exp,2} - \hat{y}_{mod}^{(k)}(\mathbf{x}_2) \\ \vdots \\ y_{exp,N_{exp}} - \hat{y}_{mod}^{(k)}(\mathbf{x}_{N_{exp}} \end{pmatrix} \in \Re^{N_{exp}}$ e

$$\mathbf{A}_k = \begin{pmatrix} f_0^{(k)}(\mathbf{x}_1) & f_1^{(k)}(\mathbf{x}_1) & \cdots & f_n^{(k)}(\mathbf{x}_1) \\ f_0^{(k)}(\mathbf{x}_2) & f_1^{(k)}(\mathbf{x}_2 & \cdots & f_n^{(k)}(\mathbf{x}_2) \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ f_0^{(k)}(\mathbf{x}_{N_{exp}} & f_1^{(k)}(\mathbf{x}_{N_{exp}}) & \cdots & f_n^{(k)}(\mathbf{x}_{N_{exp}}) \end{pmatrix}.$$

Conforme observado na seção anterior, quando o método de Newton é utilizado para a solução de problemas de mínimos quadrados, ele é comumente referenciado na literatura como método de Gauss-Newton. Definindo então a função resíduo:

$$R_j(\mathbf{a}) = y_{exp,j} - y_{mod}(\mathbf{x}_j,\mathbf{a}).$$

A função objetivo pode ser escrita como:

$$S(\mathbf{a}) = \mathbf{R}^T(\mathbf{a})\mathbf{R}(\mathbf{a}) = \sum_{j=1}^{N_{exp}} R_j^2(\mathbf{a})$$

neste caso $\nabla S(\mathbf{a}^{(k)}) = -2\mathbf{A}_k^T\mathbf{R}(\mathbf{a}^{(k)})$ e $\mathbf{H}(\mathbf{a}^{(k)}) = 2\mathbf{A}_k^T\mathbf{A}_k - 2\mathbf{Q}(\mathbf{a}^{(k)})$, em que $\mathbf{Q}(\mathbf{a}) = \sum_{j=1}^{N_{exp}} R_j(\mathbf{a})\mathbf{H}_j(\mathbf{a})$ e $\mathbf{H}_j(\mathbf{a})$ é a matriz Hessiana da função $R_j(\mathbf{a})$. Quando $R_j(\mathbf{a}^{(k)}) \to 0$, então $\mathbf{Q}(\mathbf{a}^{(k)})$ tende a zero, e as direções de busca do método de Gauss-Newton para o problema de mínimos quadrados:

$$\min_{\mathbf{d}\in\Re^n}\left\|\mathbf{R}(\mathbf{a}^{(k)}) - \mathbf{A}_k\mathbf{d}\right\|^2$$

são equivalentes às direções do método de Newton, ou seja:

$$\mathbf{d}^k = \left(\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}\mathbf{A}_k^T\mathbf{R}(\mathbf{a}^{(k)}) \approx -[\mathbf{H}(\mathbf{a}^{(k)})]^{-1}\nabla S(\mathbf{a}^{(k)}).$$

Casos Particulares de Modelos Não Lineares nos Parâmetros

1) **Modelos Exponenciais**: $y_{mod}(x,\mathbf{a},\mathbf{b}) = \sum_{i=0}^{n} a_i \exp(b_i x)$

Neste caso, se os pontos experimentais forem igualmente espaçados, isto é: $x_j = x_1 + (j-1)h$ para $j = 1,2,...,N_{exp}$, tem-se:

$$y_{mod}(x_j, \mathbf{c}, \mathbf{p}) = \sum_{i=0}^{n} c_i\, p_i^{j-1}, \text{ sendo: } c_i = a_i \exp(b_i x_1) \text{ e } p_i = \exp(b_i h)$$

Desse modo, em cada ponto experimental se tem:

$$R_j = y_{exp,j} - \sum_{i=0}^{n} c_i\, p_i^{j-1}$$

Valores preliminares dos parâmetros c_i e p_i podem ser obtidos considerando que:

$$R_j = y_{exp,j} - \sum_{i=0}^{n} c_i\, p_i^{j-1} \approx 0 \Rightarrow y_{exp,j} \cong \sum_{i=0}^{n} c_i\, p_i^{j-1},$$

considerando que $p_0, p_1, ..., p_n$ são os valores característicos de uma equação de diferenças de ordem $n+1$, linear, de coeficientes constantes e homogênea da forma: $Y_{j+n+1} + \sum_{k=0}^{n} \alpha_k Y_{j+k} = 0$. Porém, nos pontos experimentais tal equação não é satisfeita totalmente, assim define-se o resíduo:

$$\Re_j(\alpha) = y_{exp,j+n+1} + \sum_{k=0}^{n} \alpha_k\, y_{\exp,j+k} \text{ para } 1 \le j \le J_{max} = N_{exp} - (n+1)$$

Os valores de $\alpha_0, \alpha_1, ..., \alpha_n$ são determinados de modo a minimizar a função:

$$F(\alpha) = \sum_{j=1}^{J_{max}} \Re_j^2 = \sum_{j=1}^{J_{max}} \left[y_{exp,j+n+1} + \sum_{k=0}^{n} \alpha_k\, y_{exp,j+k} \right]^2$$

sendo $J_{max} = N_{exp} - (n+1) \ge 1$. Assim:

$$\frac{\partial F(\alpha)}{\partial \alpha_m} = 2 \sum_{j=1}^{J_{max}} y_{exp,j+m} \left[y_{exp,j+n+1} + \sum_{k=0}^{n} \alpha_k\, y_{exp,j+k} \right] = 0$$

ou seja:

$$\sum_{j=1}^{J_{max}} y_{exp,j+m} \left[y_{exp,j+n+1} + \sum_{k=0}^{n} \alpha_k\, y_{exp,j+k} \right] = 0 \text{ para } m = 0,1,...,n.$$

Definindo: $\mathbf{M} = \begin{pmatrix} y_{exp,1} & y_{exp,2} & y_{exp,3} & \cdots & y_{exp,n+1} \\ y_{exp,2} & y_{exp,3} & y_{exp,4} & \cdots & y_{exp,n+2} \\ y_{exp,3} & y_{exp,4} & y_{exp,5} & \cdots & y_{exp,n+3} \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ y_{exp,J_{max}} & y_{exp,J_{max}+1} & y_{exp,J_{max}+2} & \cdots & y_{exp,N_{exp}-1} \end{pmatrix}$ e $\mathbf{v} = \begin{pmatrix} y_{exp,n+2} \\ y_{exp,n+3} \\ y_{exp,n+4} \\ \vdots \\ y_{exp,N_{exp}} \end{pmatrix}$,

tem-se: $\left(\mathbf{M}^T \mathbf{M}\right) \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = -\mathbf{M}^T \mathbf{v}$. Com os valores dos coeficientes $\alpha_0, \alpha_1, ..., \alpha_n$, determinam-se as $(n+1)$ raízes do polinômio: $P_{n+1}(r) = r^{n+1} + \sum_{k=0}^{n} \alpha_k\, r^k$. Sejam essas raízes: $p_0, p_1, \cdots, p_n$, então,

em vista de $p_i = \exp(b_i h) \Rightarrow b_i = \frac{1}{h}\ln(p_i)$ para $i = 0,1,...,n$, $y_{mod}(x,\mathbf{a},\mathbf{b}) = \sum_{i=0}^{n} a_i \exp(b_i x)$. Como agora a equação do modelo é linear nos coeficientes $a_0, a_1, ..., a_n$, esses são determinados após a definição de:

$$A = \begin{pmatrix} \exp(b_0 x_1) & \exp(b_1 x_1) & \cdots & \exp(b_n x_1) \\ \exp(b_0 x_2) & \exp(b_1 x_2) & \cdots & \exp(b_n x_2) \\ \vdots & \vdots & \vdots\vdots & \vdots \\ \exp(b_0 x_{N_{exp}}) & \exp(b_1 x_{N_{exp}}) & \cdots & \exp(b_n x_{N_{exp}}) \end{pmatrix}$$

permitindo determinar: $\left[\mathbf{A}^T\mathbf{A}\right]\mathbf{a} = \mathbf{A}^T\mathbf{y}_{exp} \Rightarrow \mathbf{a} = \left[\mathbf{A}^T\mathbf{A}\right]^{-1}\left(\mathbf{A}^T\mathbf{y}_{exp}\right)$.

Como exemplo do procedimento, considera-se $n = 0$, isto é: $y_{mod}(x,a,b) = a\exp(bx)$.

Resultando em: $\mathbf{M} = \begin{pmatrix} y_{exp,1} \\ y_{exp,2} \\ y_{exp,3} \\ \vdots \\ y_{exp,N_{exp}-1} \end{pmatrix} \Rightarrow \mathbf{M}^T\mathbf{M} = \sum_{i=1}^{N_{exp}-1} y_{exp,i}^2$ e

$$\mathbf{v} = \begin{pmatrix} y_{exp,2} \\ y_{exp,3} \\ y_{exp,4} \\ \vdots \\ y_{exp,N_{exp}} \end{pmatrix} \Rightarrow \mathbf{M}^T\mathbf{v} = \sum_{i=2}^{N_{exp}} \left(y_{exp,i}\, y_{exp,i-1}\right).$$

Desse modo:

$$\alpha_0 = -\frac{\sum_{i=2}^{N_{exp}} \left(y_{exp,i}\, y_{exp,i-1}\right)}{\sum_{i=1}^{N_{exp}-1} y_{exp,i}^2} \Rightarrow p_0 = \exp(bh) = -\alpha_0 \Rightarrow b = \frac{1}{h}\ln\left(\frac{\sum_{i=2}^{N_{exp}} \left(y_{exp,i}\, y_{exp,i-1}\right)}{\sum_{i=1}^{N_{exp}-1} y_{exp,i}^2}\right)$$

e $a = \dfrac{\sum_{j=1}^{N_{exp}} \exp(bx_j)\, y_{exp,j}}{\sum_{j=1}^{N_{exp}} \exp(2bx_j)}$.

Para refinar os valores de a e b assim procede-se:

Minimiza-se a função: $S(a,b) = \sum_{j=1}^{N_{exp}} \left[y_{exp,j} - a\exp(bx_j)\right]^2$ obtendo-se:

$$\frac{\partial S(a,b)}{\partial a} = -2\sum_{j=1}^{N_{exp}} \exp(bx_j)\left[y_{exp,j} - a\exp(bx_j)\right] = 0 \Rightarrow a(b) = \frac{\sum_{j=1}^{N_{exp}} \exp(bx_j)\, y_{exp,j}}{\sum_{j=1}^{N_{exp}} \exp(2bx_j)}$$

e $\frac{\partial S(a,b)}{\partial b} = -2a\sum_{j=1}^{N_{exp}} x_j \exp(bx_j)\left[y_{exp,j} - a\exp(bx_j)\right] = 0.$

Essa última expressão, após a substituição de a em função de b, é uma função não linear apenas de b que é resolvida numericamente por um método adequado.

Muitos trabalhos na literatura utilizam a função $y_{mod}(x,a,b) = a\exp(bx)$ em sua forma logarítmica:

$$\ln[y_{mod}(x,a,b)] = Y_{mod}(x,\alpha,b) = \ln(a) + bx = \alpha + bx, \text{ sendo } \alpha = \ln(a) \Rightarrow a = \exp(\alpha).$$

Essa nova função é considerada como a equação do modelo e os valores de α e b são calculados da mesma forma que no ajuste linear, assim:

$$\begin{pmatrix} 1 & \langle x\rangle \\ \langle x\rangle & \langle x^2\rangle \end{pmatrix}\begin{pmatrix} \alpha \\ b \end{pmatrix} = \begin{pmatrix} \langle Y_{exp}\rangle \\ \langle xY_{exp}\rangle \end{pmatrix}, \text{ sendo } Y_{exp} = \begin{pmatrix} \ln(y_{exp,1}) \\ \ln(y_{exp,2}) \\ \vdots \\ \ln(y_{exp,N_{exp}}) \end{pmatrix}.$$

Esse procedimento é denominado erroneamente de *linearização* e seu emprego não é recomendável quando se deseja um bom ajuste aos dados. Porém, os valores de a e b assim estimados podem ser utilizados como valores iniciais para o procedimento mais rigoroso.

2) **Modelo Hiperbólico**: $y_{mod}(x,a,b) = \frac{a}{1+bx}$

De forma semelhante da feita com o modelo exponencial acima, este modelo pode também ser expresso na forma: $Y_{mod}(x,\alpha,\beta) = \frac{1}{y_{mod}(x,a,b)} = \alpha + \beta\, x$ sendo:

$$\alpha = \frac{1}{a} \Rightarrow a = \frac{1}{\alpha} \text{ e } \beta = \frac{b}{a} \Rightarrow b = \frac{\beta}{\alpha}.$$

Os valores de α e β são calculados da mesma forma que no ajuste linear, assim:

$$\begin{pmatrix} 1 & \langle x\rangle \\ \langle x\rangle & \langle x^2\rangle \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \langle Y_{exp}\rangle \\ \langle xY_{exp}\rangle \end{pmatrix}, \text{ sendo } \mathbf{Y}_{exp} = \begin{pmatrix} \frac{1}{y_{exp,1}} \\ \frac{1}{y_{exp,2}} \\ \vdots \\ \frac{1}{y_{exp,N_{exp}}} \end{pmatrix}.$$

Neste caso também vale as observações feitas acima. Para um ajuste que minimize de fato a soma dos quadrados dos erros deve-se:
Minimizar a função: $S(a,b) = \sum_{j=1}^{N_{exp}} \left[y_{exp,j} - \frac{a}{1+bx_j} \right]^2$, obtendo-se:

$$\frac{\partial S(a,b)}{\partial a} = -2\sum_{j=1}^{N_{exp}} \frac{1}{1+bx_j}\left[y_{exp,j} - \frac{a}{1+bx_j} \right] = 0 \Rightarrow a(b) = \frac{\sum_{j=1}^{N_{exp}} \frac{y_{\exp,j}}{1+bx_j}}{\sum_{j=1}^{N_{exp}} \left(\frac{1}{1+bx_j} \right)^2}$$

e $\frac{\partial S(a,b)}{\partial b} = 2a\sum_{j=1}^{N_{exp}} \frac{x_j}{(1+bx_j)^2}\left[y_{\exp,j} - \frac{a}{1+bx_j} \right] = 0$

Essa última expressão, após a substituição de a em função de b, é uma função não linear apenas de b que é resolvida numericamente por um método adequado.

Procedimento semelhante pode também ser aplicado à função:
$y_{mod}(x,a,b) = \frac{ax}{1+bx} \Rightarrow \frac{1}{y_{mod}(x,a,b)} = Y_{mod}(X,\alpha,\beta) = \alpha + \beta\, X$, sendo: $X = \frac{1}{x}$, $\alpha = \frac{b}{a}$ e $\beta = \frac{1}{a}$.

3) **Modelo Geométrico**: $y_{mod}(x,a,b) = ax^b$

De forma semelhante da feita com o modelo exponencial acima, este modelo pode também ser expresso na forma:

$$Y_{mod}(X,\alpha,b) = \ln(y_{mod}(x,a,b)) = \alpha + bX, \text{ sendo } X = \ln(x) \text{ e } \alpha = \ln(a) \Rightarrow a = \exp(\alpha).$$

Os valores de α e b são calculados da mesma forma que no ajuste linear, assim:

$$\begin{pmatrix} 1 & \langle X\rangle \\ \langle X\rangle & \langle X^2\rangle \end{pmatrix}\begin{pmatrix} \alpha \\ b \end{pmatrix} = \begin{pmatrix} \langle Y_{exp}\rangle \\ \langle xY_{exp}\rangle \end{pmatrix}, \text{ sendo } Y_{exp} = \begin{pmatrix} \ln(y_{exp1}) \\ \ln(y_{exp,2}) \\ \vdots \\ \ln(y_{exp,N_{exp}}) \end{pmatrix}.$$

Neste caso também vale as observações anteriores. Para um ajuste que minimize de fato a soma dos quadrados dos erros deve-se:

Minimizar a função: $S(a,b) = \sum_{j=1}^{N_{exp}} \left[y_{exp,j} - a x_j^b\right]^2$, obtendo-se:

$$\frac{\partial S(a,b)}{\partial a} = -2\sum_{j=1}^{N_{exp}} x_j^b \left[y_{exp,j} - a x_j^b\right] = 0 \Rightarrow a(b) = \frac{\sum_{j=1}^{N_{exp}} x_j^b y_{exp,j}}{\sum_{j=1}^{N_{exp}} x_j^{2b}}$$

e $\frac{\partial S(a,b)}{\partial b} = -2a\sum_{j=1}^{N_{exp}} x_j^b \ln(x_j)\left[y_{exp,j} - a x_j^b\right] = 0.$

Essa última expressão, após a substituição de a em função de b, é uma função não linear apenas de b que é resolvida numericamente por um método adequado.

8.5 Problemas Propostos

Problema 8.1 Determine a concavidade ou convexidade, a partir de suas definições, das seguintes funções:

(a) $f(x_1,x_2) = (x_1-2)^2 + 3(x_2+1)^2$ no domínio: $x_1,x_2 \geq 0$;
(b) $f(x_1,x_2,x_3) = 2x_1^2 + x_2^2 - 3x_3^2$ no domínio: $x_1,x_2,x_3 \geq 0$.

Problema 8.2 Determine a localização e a natureza dos pontos estacionários das funções abaixo, determine também (em cada caso) o máximo e o mínimo globais:

(a) $f(x) = \frac{2x}{1+x^2}$ para $x \geq 0$;
(b) $f(x) = \frac{sen(x)}{x}$ para $\pi \leq x \leq \pi$;
(c) $f(x) = e^{-x^2} sen(x)$ para $0 \leq x \leq \pi$;
(d) $f(x_1,x_2) = x_1^3 - 3x_1x_2 + 3x_2^2$;
(e) $f(x_1,x_2) = -x_1^2 - \frac{1}{2}x_2^2 + x_1x_2 + x_1$ para $x_1,x_2 \geq 0$.

Problema 8.3 Resolva, utilizando multiplicadores de Lagrange, cada um dos problemas abaixo:

(a) Maximize: $f(x_1,x_2) = x_1x_2$ tal que: $x_1 + 2x_2 = 4$;
(b) Minimize: $f(x_1,x_2) = (x_1-2)^2 + 2(x_2-1)^2 + (x_3-3)^2$ tal que:

$$2x_1 + x_2 + 2x_3 \geq 4 \text{ e } x_1^2 + 2x_2^2 + 3x_3^2 \geq 48;$$

(c) Minimize: $f(x_1,x_2) = (x_1-2)^2 + 2(x_2-1)^2 + (x_3-3)^2$ tal que:

$$2x_1 + x_2 + 2x_3 \leq 4 \text{ e } x_1^2 + 2x_2^2 + 3x_3^2 \leq 48;$$

(d) Minimize: $f(x_1,x_2) = x_1^2 + x_2^2$ tal que: $(x_1-1)^3 - x_2^2 = 0$.

Problema 8.4 Se as coordenadas x_1 e x_2 estão relacionadas por: $2x_1 + x_2 = 1$, ache os pontos sobre a elipsóide: $2x_1^2 + x_2^2 + x_3^2 = 1$ que se encontram, respectivamente, mais próximo e mais afastado da origem.

Problema 8.5 Determine as dimensões do paralelepípedo, cuja diagonal tem um comprimento d, que apresenta o maior volume.

Problema 8.6 Teste as condições necessárias e suficientes do problema abaixo.

$\min_{x \in \Re^2} S(x) = x_1 x_2$

sujeito a: $g_1(x) = x_1^2 + x_2^2 - 25 \leq 0$

Problema 8.7 Em um reator químico é conduzida uma reação química irreversível de segunda ordem, o processo é em batelada. O balanço de massa do reagente é descrito pela equação diferencial:

$$\frac{dc(t)}{dt} = -k\,[c(t)]^2 \;\text{ com }\; c(0) = c_0$$

A variação da concentração do reagente com o tempo é medida construindo-se a tabela:

t (min)	1	2	3	4	5	7	10	12	15	20	25
$c \times 100$ (mol/L)	4,049	3,086	2,604	2,222	1,912	1,524	1,142	0,980	0,741	0,649	0,521

Baseado nestes dados estime os valores de k e de c_0.

Dica: A solução da EDO é: $c(t) = \frac{c_0}{1+k\,c_0\,t}$ considere: $a = c_0$ e $b = k\,c_0$ e determine os parâmetros a e b como indicado na Seção 8.4 para modelo hiperbólico.

Problema 8.8 A intensidade de radiação de uma fonte radioativa é expressa por: $I(t) = I_0 e^{-\alpha t}$. Determine os valores de I_0 e de α que melhor ajustem os dados experimentais abaixo:

t	0,2	0,3	0,4	0,5	0,6	0,7	0,8
$I(t)$	3,16	2,38	1,75	1,34	1,00	0,74	0,56

Dica: note que os pontos estão igualmente espaçados, considere então que: $I_{mod}(t_{i+1}) = p\,I_{mod}(t_i)$ sendo $t_i = 0,2 + 0,1i$ para $i = 0,1,...,7$.

Problema 8.9 y é uma função de x dada pela tabela abaixo, sabe-se que esta dependência é expressa por: $y(x) = A\,e^{-\alpha x} + B\,e^{-\beta x}$. Determine os valores de A, B, α e β.

x	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1
$y(x)$	2,31604	2,02877	1,78030	1,56513	1,37854	1,21651	1,07561	0,95289

Dica: note que os pontos estão igualmente espaçados, considere então que: $y_{mod}(t_{i+2}) + b\,y_{mod}(t_{i+1}) + c\,y_{mod}(t_i) = 0$ sendo $t_i = 0,4 + 0,1i$ para $i = 0,1,...,7$. Os valores de b e c são determinados de modo a minimizar a função:

$\sum_{i=0}^{5} [y_{mod}(t_{i+2}) + b\,y_{mod}(t_{i+1}) + c\,y_{mod}(t_i)]^2 = 0$ com os valores de b e c determinam-se as raízes r_1 e r_2 de $p^2 + b\,p + c = 0$, sendo $r_1 = e^{-0,1\alpha} \Rightarrow \alpha = -10\ln(r_1)$ e $r_2 = e^{-0,1\beta} \Rightarrow \beta = -10\ln(r_2)$

Problema 8.10 y é uma função de x dada pela tabela abaixo, sabe-se que esta dependência é expressa por: $y(x) = Ce^{-ax} sen(bx)$. Determine os valores de C, a e b.

x	0,0	0,2	0,4	0,6	0,8	1,0	1,2	1,4	1,6
$y(x)$	0,00000	0,15398	0,18417	0,16156	0,12301	0,08551	0,05537	0,03362	0,01909

Dica: note que os pontos estão igualmente espaçados, considere então que: $y_{mod}(t_{i+2}) + b\,y_{mod}(t_{i+1}) + c\,y_{mod}(t_i) = 0$ sendo $t_i = 0,2i$ para $i = 0,1,...,8$. Tendo em vista que $sen(bx) = \frac{e^{bxi} - e^{-bxi}}{2i} \Rightarrow e^{-ax} sen(bx) = \frac{e^{(-a+bx)i} - e^{(-a-bx)i}}{2i}$, pode-se assim interpretar como se o polinômio característico associado apresenta um par de raiz complexa conjugada. Após as raízes serem determinadas os valores iniciais dos parâmetros seriam: $a = \frac{1}{h}\ln|\Re e(r_0)| = 5\ln|\Re e(r_0)|$ e $b = \frac{1}{h}|\arg(r_0)| = 5|\arg(r_0)|$

Problema 8.11 y é uma função de x dada pela tabela abaixo, sabe-se que esta dependência é expressa por: $y(x) = Ae^{-\alpha x} + B$. Determine os valores de A, B e α.

x	1,0	1,2	1,4	1,6	1,8	2,0	2,2	2,4	2,6	2,8	3,0
$y(x)$	3,00767	2,79720	2,61553	2,45874	2,32340	2,20659	2,10576	2,01874	1,94363	1,87880	1,82284

Dica: note que os pontos estão igualmente espaçados, considere então que: $y_{mod}(t_{i+2}) - (1+b)y_{mod}(t_{i+1}) + b\,y_{mod}(t_i) = 0$ sendo $t_i = 1+0,2i$ para $i = 0,1,...,10$. Com o valor de b determinam-se as raízes de $p^2 - (1+b)p + b = (p-1)(p-b) = 0$ estas raízes são: $r_1 = 1$ e $r_2 = b = e^{-0,2\alpha} \Rightarrow \alpha = -5\ln(b)$. Calculam-se então A e B por regressão linear.

Problema 8.12 Através de fotografias estroboscópicas de pequenas bolhas de ar é possível medir o perfil de velocidade próxima à parede de um tubo no qual escoa um fluido. Com um número de Reynolds de 1200 e com um tubo de 1 polegada de diâmetro interno os seguintes pontos experimentais são obtidos:

y (distância à parede) cm	u (velocidade) cm/s	y (distância à parede) cm	u (velocidade) cm/s
0,003	0,03	0,056	0,85
0,021	0,32	0,061	0,92
0,025	0,30	0,070	1,05
0,025	0,33	0,078	1,117
0,037	0,57	0,085	1,32
0,043	0,66	0,092	1,38
0,049	0,74	0,106	1,57
0,053	0,80	0,113	1,65
0,055	0,84		

A função que melhor ajusta o perfil de velocidade é: $u(y) = py + qy^2$, baseado nos dados acima determine os valores de p e q.

Problema 8.13 Os coeficientes de transferência de calor em trocadores de calor são adequadamente modelados por expressão do tipo: $Nu = \alpha Re^{\beta}\,\mathrm{Pr}^{\gamma} r^{\delta}$
em que Nu, Re e Pr são, respectivamente, os números de Nusselt, Reynolds e Prandtl e r é a razão entre a viscosidade, a temperatura média do fluido e a temperatura da parede; α, β, γ e δ são constantes. Os seguintes dados experimentais estão disponíveis:

Nu	Re	Pr	r
277	49000	2,30	0,947
348	68600	2,28	0,954
421	84800	2,27	0,959
223	34200	2,32	0,943
177	22900	2,36	0,936
114,8	1321	246	0,592
95,9	931	247	0,583
68,3	518	251	0,579
49,1	346	273	0,290
56,0	122,9	1518	0,294
39,9	54,0	1590	0,279
47,0	84,6	1521	0,267
94,2	1249	107,4	0,724
99,9	1021	186	0,612
83,1	465	414	0,512
35,9	54,8	1302	0,273

Estimar os valores de α, β, γ e δ que melhor ajustam os pontos acima.

Problema 8.14 Encontre o mínimo das seguintes funções objetivo usando os métodos diretos e indiretos descritos nas Seções 8.2 e 8.3 e compare os resultados em termos do número de funções objetivo avaliadas por cada método:

a) $S(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$
b) $S(x) = [1,5 - x_1(1 - x_2)^2 + [2,25 - x_1(1 - x_2^2)]^2 + [2,625 - x_1(1 - x_2^3)]^2$
c) $S(x) = 4x_1^2 - 2x_1x_2 + x_2^2$
d) $S(x) = exp(x_1)(4x_1^2 + 2x_2^2 + 4x_1x_2 + 2x_2 + 1)$
e) $S(x) = 4(x_1 - 5)^2 + (x_2 - 6)^2$
f) $S(x) = x_1^2 - 5x_1 + 3x_2^2 + 3$
g) $S(x) = (x_1 - 2)^2 + (x_2 - 1)^2$.

A. Elementos de Álgebra Linear

A.1 Conceitos Básicos

Os cálculos/operações assim como conceitos envolvendo matrizes e vetores constituem a base dos métodos numéricos que tratam da solução de sistemas lineares e não lineares de equações algébricas ou diferenciais. A representação desses sistemas em termos matriciais/vetoriais é extremamente mais compacta e é corrente na literatura técnica. Como esses conceitos são utilizados na resolução de problemas típicos da Engenharia Química, os elementos de matrizes e vetores são em princípio números ou variáveis reais a não ser quando explicitamente especificados como complexos.

Uma matriz é um arranjo retangular de números em m linhas e n colunas, $(m \times n)$, sendo representada neste texto como $\mathbf{A}$ (letras maiúsculas em negrito) pertencente a $\Re^{m\times n}$, expresso por $\mathbf{A} \in \Re^{m\times n}$. O elemento da linha i e coluna j de $\mathbf{A}$ é representado por a_{ij} (correspondente letra minúscula com o sub-índice ij) ou $\mathbf{A}_{ij}$.

A matriz completa é geralmente escrita na forma: $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$ ou, em forma mais compacta, por: $\mathbf{A} = (a_{ij})$ com $i = 1, \cdots, m$ e $j = 1, \cdots, n$.

Se duas matrizes $\mathbf{A}$ e $\mathbf{B}$ apresentam o mesmo número de linhas e o mesmo número de colunas são ditas do mesmo *tipo*.

Se $\mathbf{A} = (a_{ij})$ é tal que $a_{ij} = 0$ para todo i e j então a matriz $\mathbf{A}$ é dita nula e é representada por $\mathbf{0}$.

Se $m = n$ a matriz $\mathbf{A}$ é dita *quadrada*.

Se $m = n$ e $a_{ij} = a_{ji}$ a matriz $\mathbf{A}$ é dita *simétrica*.

Se $n = 1$ tem-se um *vetor coluna*, ou simplesmente *vetor*, designado por $\mathbf{v}$ (letra minúscula em negrito) e representado por $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \in \Re^m$.

Se $m=1$ tem-se um *vetor linha* designado por $\mathbf{v}^T$ (letra minúscula em negrito com sobrescrito T de transposto) e representado por $\mathbf{v}^T=\begin{pmatrix}v_1 & v_2 & \cdots & v_n\end{pmatrix}\in\Re^{1\times n}$.

Se $m=n=1$ tem-se um *escalar* (real) α (letra minúscula grega), $\alpha\in\Re$.

A matriz $\mathbf{A}\in Re^{m\times n}$ pode ser *particionada* por:

(a) Colunas na forma:

$$\mathbf{A}=\begin{pmatrix}\mathbf{a}^{\langle 1\rangle} & \mathbf{a}^{\langle 2\rangle} & \cdots & \mathbf{a}^{\langle n\rangle}\end{pmatrix},\ \text{em que}\ \mathbf{a}^{\langle j\rangle}=\begin{pmatrix}a_{1j}\\ a_{2j}\\ \vdots\\ a_{mj}\end{pmatrix},\ \text{para}\ j=1,\cdots,n,\ \text{são os vetores coluna}$$

da matriz $\mathbf{A}$.

(b) Linhas na forma:

$$\mathbf{A}=\begin{pmatrix}\mathbf{a}_1^T\\ \mathbf{a}_2^T\\ \vdots\\ \mathbf{a}_m^T\end{pmatrix}\ \text{em que}\ \mathbf{a}_i^T=\begin{pmatrix}a_{i1} & a_{i2} & \cdots & a_{in}\end{pmatrix},\ \text{para}\ i=1,\cdots,m,\ \text{são os vetores linha da}$$

matriz $\mathbf{A}$.

A.2 Operações entre Matrizes

As operações de adição ou subtração são definidas apenas para matrizes do mesmo tipo, assim se $\mathbf{A}$ e $\mathbf{B}$ são matrizes $(m\times n)$ então a matriz $\mathbf{C}$, também $(m\times n)$, *soma* ou *subtração* de $\mathbf{A}$ com $\mathbf{B}$, representada por $\mathbf{C}=\mathbf{A}\pm\mathbf{B}$, tem como termo geral $c_{ij}=a_{ij}\pm b_{ij}$ para $i=1,\cdots,m$ e $j=1,\cdots,n$.

Se α é um escalar real, então a matriz $\alpha\mathbf{A}$ é uma matriz cujo termo geral é αa_{ij}.

A operação de multiplicação de matrizes está intimamente relacionada a transformações de coordenadas. Assim sejam as seguintes *transformações lineares*:

$$z_i=\sum_{j=1}^{n}a_{ij}y_j\ \text{para}\ i=1,\cdots,m\ \text{e}\ y_j=\sum_{k=1}^{p}b_{jk}x_k\ \text{para}\ j=1,\cdots,n.$$

Expressando z_i em termos de x_k, obtém-se

$$z_i=\sum_{j=1}^{n}a_{ij}\left(\sum_{k=1}^{p}b_{jk}x_k\right)=\sum_{k=1}^{p}\left(\sum_{j=1}^{n}a_{ij}b_{jk}\right)x_k.$$

Definindo $c_{ik}=\sum_{j=1}^{n}a_{ij}b_{jk}$, resulta $z_i=\sum_{k=1}^{p}c_{ik}x_k$, sugerindo a definição da matriz $\mathbf{C}=\mathbf{A}\,\mathbf{B}$ sendo $\mathbf{A}(m,n)$, $\mathbf{B}(n,p)$ e $\mathbf{C}(m,p)$, cujo termo geral é $c_{ik}=\sum_{j=1}^{n}a_{ij}b_{jk}$ para $i=1,\cdots,m$ e $k=1,\cdots,p$.

Verificando-se assim que a operação $\mathbf{A}\,\mathbf{B}$ só é definida se o número de colunas de $\mathbf{A}$ (primeira parcela do produto) for igual ao número de linhas de $\mathbf{B}$ (segunda parcela do produto). É importante ressaltar que a lei de comutatividade não é satisfeita pelo produto entre matrizes, mesmo que $\mathbf{B}\,\mathbf{A}$ seja definida, isto é se $m=p$ e mesmo que $\mathbf{B}\,\mathbf{A}$ seja do mesmo tipo que $\mathbf{A}\,\mathbf{B}$, o que só ocorrerá se $m=p=n$ (isto é ambas as matrizes são quadradas e de mesma dimensão), assim de uma forma geral tem-se $\mathbf{A}\,\mathbf{B}\neq\mathbf{B}\,\mathbf{A}$.

Se a primeira parcela do produto for um vetor linha $\mathbf{u}^T(1,n)$ e a segunda parcela for um vetor coluna $\mathbf{v}(n,1)$ então o produto é um escalar $\mathbf{u}^T\,\mathbf{v}=\mathbf{u}\cdot\mathbf{v}=\sum_{j=1}^{n}u_jv_j$, que é comutável, isto é: $\mathbf{u}^T\,\mathbf{v}=\mathbf{v}^T\,\mathbf{u}$. Este produto é chamado de *produto escalar* de dois vetores.

Se a primeira parcela do produto for uma matriz $\mathbf{A}(m,n)$ e a segunda parcela for um vetor $\mathbf{v}(n,1)$ então o produto $\mathbf{A}\,\mathbf{v}$ é um vetor $\mathbf{u}(m,1)$ cujo termo geral é $u_i = \sum_{j=1}^{n} a_{ij} v_j$ para $i = 1, \cdots, m$. Este produto pode ser efetuado de duas formas distintas:

(a) Método **ij**. Considerando a partição por linhas da matriz $\mathbf{A}$, $\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix}$, então $\mathbf{u} = \mathbf{A}\,\mathbf{v} = \begin{pmatrix} \mathbf{a}_1^T\,\mathbf{v} \\ \mathbf{a}_2^T\,\mathbf{v} \\ \vdots \\ \mathbf{a}_m^T\,\mathbf{v} \end{pmatrix}$, cujo termo geral é $u_i = \mathbf{a}_i^T\,\mathbf{v}$, que é o produto escalar dos elementos da linha i da matriz $\mathbf{A}$ pelo vetor $\mathbf{v}$.

(b) Método **ji**. Considerando a partição por colunas da matriz $\mathbf{A}$, $\mathbf{A} = \begin{pmatrix} \mathbf{a}^{\langle 1\rangle} & \mathbf{a}^{\langle 2\rangle} & \cdots & \mathbf{a}^{\langle n\rangle} \end{pmatrix}$, então $\mathbf{u} = \mathbf{A}\,\mathbf{v} = \begin{pmatrix} \mathbf{a}^{\langle 1\rangle} & \mathbf{a}^{\langle 2\rangle} & \cdots & \mathbf{a}^{\langle n\rangle} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \sum_{i=1}^{n} v_i\,\mathbf{a}^{\langle i\rangle}$, isto é, o vetor $\mathbf{u}$ é uma *combinação linear* dos vetores coluna de $\mathbf{A}$, sendo os coeficientes desta combinação os elementos do vetor $\mathbf{v}$.

A operação de *transposição* de uma matriz $\mathbf{A}(m,n)$ consiste em trocar as linhas pelas colunas de $\mathbf{A}$, esta nova matriz é chamada de matriz *transposta* de $\mathbf{A}$, representada por $\mathbf{A}^T$ e é uma matriz (n,m) cujo termo da linha j e coluna i é $a_{ji}^T = a_{ij}$ para $j = 1, \cdots, n$ e $i = 1, \cdots, m$. Se a matriz $\mathbf{A}$ for simétrica então $\mathbf{A} = \mathbf{A}^T$.

A seguir, descrevem-se propriedades que se aplicam **apenas** a matrizes quadradas (n,n), a vetores coluna $(n,1)$ e a vetores linha $(1,n)$.

Define-se como *matriz identidade* a matriz $\mathbf{I}$ cujo elemento geral é $(\mathbf{I})_{ij} = \delta_{ij} = \begin{cases} 1 \text{ se } i = j \\ 0 \text{ se } i \neq j \end{cases}$, em que δ_{ij} é o *delta de Krönecker*. A matriz identidade é uma matriz diagonal em que todos os termos são unitários $\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \text{diag}\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$.

Sendo uma *matriz diagonal* uma matriz quadrada em que somente os elementos da diagonal não são nulos, geralmente uma matriz diagonal $\mathbf{D} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$, é representada na forma mais compacta $\mathbf{D} = \text{diag}\begin{pmatrix} d_1 & d_2 & \cdots & d_n \end{pmatrix}$. Toda matriz diagonal, em vista de $a_{ij} = 0$ se $i \neq j$, é também simétrica.

Uma propriedade importante da matriz identidade é $\mathbf{I}\,\mathbf{A} = \mathbf{A}\,\mathbf{I}$, isto é, a matriz identidade pré-multiplicada ou pós-multiplicada por qualquer matriz quadrada de mesma dimensão não altera o valor de elemento algum desta matriz. Algumas vezes, para evitar ambiguidades, representa-se a matriz identidade de dimensão n por $\mathbf{I}_n$.

Uma matriz diagonal é um caso particular de matrizes ditas *esparsas*, que são matrizes que apresentam um grande número de elementos nulos, sendo os elementos não nulos mais a exceção do que a regra. Algumas destas matrizes são apresentadas a seguir:

1. Matrizes tri-diagonais: são matrizes que apresentam apenas os elementos da diagonal, os elementos imediatamente sobre a diagonal e os elementos imediatamente sob a diagonal não nulos, sendo os demais nulos, assim se $\mathbf{A}$ for uma matriz tri-diagonal então
$$a_{i,j} = \begin{cases} \neq 0 \text{ se } i = j \text{ ou } i = j+1 \text{ ou } i = j-1 \\ = 0 \text{ em qualquer outro caso} \end{cases} .$$
2. Matrizes bi-diagonais: são matrizes que apresentam apenas os elementos da diagonal, os elementos imediatamente sobre a diagonal ou os elementos imediatamente sob a diagonal não nulos, no primeiro caso diz-se que a matriz é *bi-diagonal superior* e no segundo caso *bi-diagonal inferior*.
3. Matrizes triangulares: são matrizes que apresentam todos os elementos sob a diagonal nulos (*matriz triangular superior*) ou todos os elementos sobre a diagonal nulos (*matriz triangular inferior*). Matrizes triangulares superiores são representadas por $\mathbf{U}$ (*upper*) e matrizes triangulares inferiores são representadas por $\mathbf{L}$ (*lower*).

O *traço* de uma matriz quadrada $\mathbf{A}$ é a soma dos elementos de sua diagonal, assim
$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Uma matriz quadrada $\mathbf{A}$ é dita *positiva definida* se $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} > 0$ para todo vetor $\mathbf{x} \neq \mathbf{0}$, caso $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} \geq 0$ a matriz $\mathbf{A}$ é dita *positiva semi-definida*. Uma matriz quadrada $\mathbf{A}$ é dita *negativa definida* se $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} < 0$ para todo vetor $\mathbf{x} \neq \mathbf{0}$, caso $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} \leq 0$ a matriz $\mathbf{A}$ é dita *negativa semi-definida*. Uma matriz quadrada $\mathbf{A}$ é dita *não definida* se $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} > 0$ para algum vetor $\mathbf{x} \neq \mathbf{0}$ e ao mesmo tempo $\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} < 0$ para algum outro vetor $\mathbf{x} \neq \mathbf{0}$.

O *determinante* de uma matriz $\mathbf{A}$ é um escalar obtido através da soma de todos os produtos possíveis envolvendo um elemento de cada linha e cada coluna da matriz, com o sinal positivo ou negativo conforme o número de permutações dos índices seja par ou ímpar. Sua obtenção e sua representação, apesar de ser um dos conceitos mais preliminares envolvendo matrizes, não são tarefas triviais e o conceito de determinante é utilizado nestas notas apenas como base de outras propriedades de matrizes quadradas. Assim, o determinante de $\mathbf{A}$ designado por $\det(\mathbf{A})$ pode ser representado por: $\det(\mathbf{A}) = \sum \pm a_{1,i_1} a_{2,i_2} \cdots a_{n,i_n}$, ou então através do conceito de cofator do elemento ij da matriz $\mathbf{A}$ (representado por A_{ij}) que é o determinante da matriz obtida cancelando a linha i e a coluna j da matriz $\mathbf{A}$ com o sinal mais ou menos conforme $i+j$ seja par ou ímpar, ou seja $A_{ij} = (-1)^{i+j}\det(\Lambda_{ij})$ em que Λ_{ij} é a matriz quadrada $(n-1, n-1)$ obtida pela eliminação a linha i e a coluna j da matriz $\mathbf{A}$. Tem-se então:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij} A_{ij} \qquad\qquad \det(\mathbf{A}) = \sum_{i=1}^{n} a_{ij} A_{ij}$$

Expansão do determinante pela linha i Expansão do determinante pela coluna j

Estas expansões apresentam as propriedades

$$\begin{cases} \sum_{j=1}^{n} a_{ij} A_{kj} = 0 \text{ se } k \neq i, \text{ equivalente a afirmar que a matriz } \mathbf{A} \text{ apresenta duas linhas iguais} \\ \sum_{i=1}^{n} a_{ij} A_{ik} = 0 \text{ se } k \neq j, \text{ equivalente a afirmar que a matriz } \mathbf{A} \text{ apresenta duas colunas iguais} \end{cases}$$

Na prática, entretanto, é praticamente impossível calcular o determinante de matrizes através destas regras gerais por envolver um número muito grande de termos (na realidade $n!$ assim mesmo com matrizes relativamente pequenas tais como $n = 10$ tem-se 3 milhões de termos). Entretanto, para os propósitos deste apêndice apenas as seguintes propriedades de determinante de matrizes são pertinentes:

1. O determinante de uma matriz $\mathbf{A}$ mantém-se inalterado ao somar a todos os elementos de qualquer linha (ou coluna) os correspondentes elementos de uma outra linha (ou coluna) multiplicados pela mesma constante α.
2. Se a_{ij} for o único elemento não nulo da linha i ou da coluna j então $\det(\mathbf{A}) = a_{ij}A_{ij}$.
3. Se $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, então $\det(\mathbf{A}) = ad - bc$.
4. $\det(\mathbf{A}\,\mathbf{B}) = \det(\mathbf{B}\,\mathbf{A}) = \det(\mathbf{A})\det(\mathbf{B})$.
5. $\det(\mathbf{A}^T) = \det(\mathbf{A})$.

Como corolário da propriedade 1. tem-se que se $\det(\mathbf{A}) = 0$, então $\mathbf{A}$ apresenta duas linhas (ou colunas) proporcionais entre si, ou ainda, de uma forma mais geral, pode-se afirmar que uma linha (ou coluna) de $\mathbf{A}$ pode ser escrita como combinação linear de alguma ou algumas linhas (ou colunas) da mesma matriz. Da propriedade 2. demonstra-se que se $\mathbf{A}$ for uma matriz triangular, então $\det(\mathbf{A})$ é simplesmente o produto dos elementos de sua diagonal (o mesmo valendo para matrizes bi-diagonais por serem também matrizes triangulares).

Se $\det(\mathbf{A}) = 0$ diz-se que a matriz $\mathbf{A}$ é *singular*, e caso $\det(\mathbf{A}) \neq 0$ diz-se que a matriz $\mathbf{A}$ é *regular* ou *não singular*.

A *matriz adjunta* de uma matriz $\mathbf{A}$ é a matriz transposta da matriz obtida substituindo cada elemento da matriz $\mathbf{A}$ pelo seu correspondente cofator, isto é, se $\text{adj}(\mathbf{A})$ for a matriz adjunta de $\mathbf{A}$, então o elemento da linha i e coluna j de $\text{adj}(\mathbf{A})$, é A_{ij}. A propriedade mais importante da matriz adjunta diz respeito aos produtos: $\mathbf{P} = \mathbf{A}\ \text{adj}(\mathbf{A})$ e $\mathbf{Q} = \text{adj}(\mathbf{A})\ \mathbf{A}$. O primeiro produto tem com termo geral $p_{ij} = \sum_{k=1}^{n} a_{ik}\hat{a}_{kj} = \sum_{k=1}^{n} a_{ik}A_{kj} = \det(\mathbf{A})\delta_{ij}$ e o termo geral do segundo produto é $q_{ij} = \sum_{k=1}^{n} \hat{a}_{ik}a_{kj} = \sum_{k=1}^{n} A_{ki}a_{kj} = \det(\mathbf{A})\delta_{ij}$, permitindo concluir que $\mathbf{A}\ \text{adj}(\mathbf{A}) = \text{adj}(\mathbf{A})\ \mathbf{A} = \det(\mathbf{A})\mathbf{I}$.

Desse modo se $\det(\mathbf{A}) \neq 0$ define-se $\mathbf{A}^{-1} = \dfrac{\text{adj}(\mathbf{A})}{\det(\mathbf{A})}$ como sendo a *inversa* de $\mathbf{A}$ que apresenta a propriedade $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{I}$, enfatizando-se que $\mathbf{A}^{-1}$ só existe se $\det(\mathbf{A}) \neq 0$. Além disto, tem-se $\det(\mathbf{A}^{-1}) = \dfrac{1}{\det(\mathbf{A})}$.

Se $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, cujos cofatores são $\begin{cases} A_{11} = d,\ A_{12} = -c \\ A_{21} = -b,\ A_{22} = a \end{cases} \Rightarrow \text{adj}(\mathbf{A}) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Verificando-se que

$$\mathbf{A}\ \text{adj}(\mathbf{A}) = \text{adj}(\mathbf{A})\ \mathbf{A} = (ad - bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = (ad - bc)\mathbf{I} \Rightarrow \mathbf{A}^{-1} = \frac{1}{(ad - bc)}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Concluindo-se que para determinar a inversa de uma matriz (2×2) basta trocar os elementos da diagonal principal, trocar o sinal dos elementos da diagonal secundária e dividir a matriz resultante pelo determinante da matriz original.

Se $\mathbf{A}^{-1} = \mathbf{A}^T$ a matriz $\mathbf{A}$ é chamada de *matriz ortogonal* e, neste caso, como $\det(\mathbf{A}^T) = \det(\mathbf{A})$ e $\det(\mathbf{A}^{-1}) = \dfrac{1}{\det(\mathbf{A})}$, resulta em $\dfrac{1}{\det(\mathbf{A})} = \det(\mathbf{A}) \Rightarrow [\det(\mathbf{A})]^2 = 1$, logo $\det(\mathbf{A}) = +1$ ou -1.

■ **Exemplo A.1** Um exemplo ilustrativo do emprego de matrizes ortogonais é o relativo à transformação linear em decorrência da rotação dos eixos de coordenadas, representado na Figura A.1.

Verifica-se que as coordenadas originais são $\begin{cases} x_1 = r\cos(\alpha) \\ x_2 = r\,\text{sen}(\alpha) \end{cases}$, e os valores das novas coordenadas são

$$\begin{cases} y_1 = r\cos(\theta - \alpha) = r\cos(\alpha)\cos(\theta) + r\,\text{sen}(\alpha)\,\text{sen}(\theta) = x_1\cos(\theta) + x_2\,\text{sen}(\theta) \\ y_2 = -r\,\text{sen}(\theta - \alpha) = -r\cos(\alpha)\,\text{sen}(\theta) + r\,\text{sen}(\alpha)\cos(\theta) = -x_1\,\text{sen}(\theta) + x_2\cos(\theta) \end{cases}.$$

Representando esta transformação em termos matriciais:

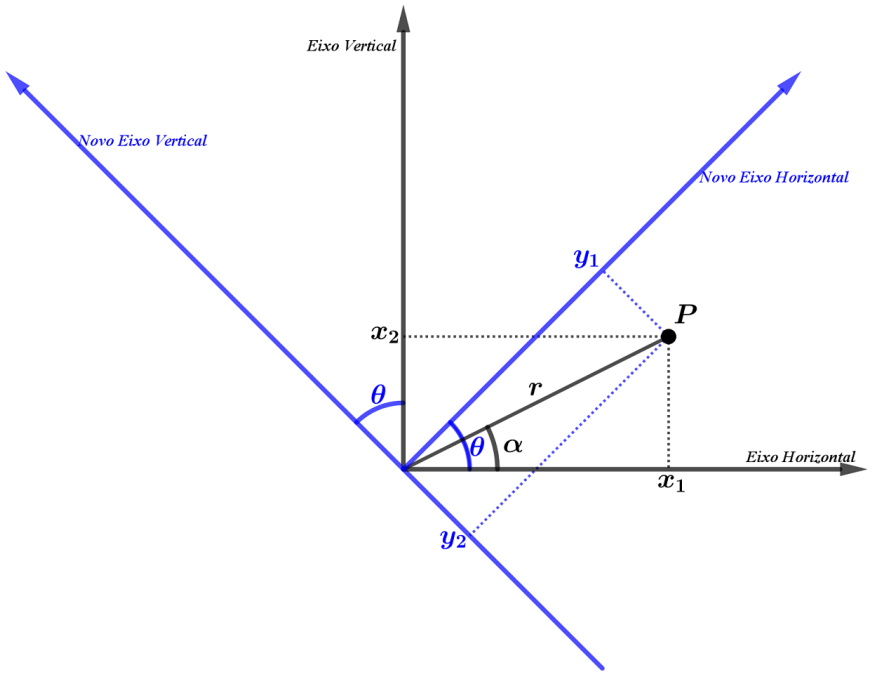


Figura A.1: Mudança de coordenadas por rotação dos eixos.

$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \text{sen}(\theta) \\ -\text{sen}(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, identificando a matriz transformação:

$\mathbf{M} = \begin{pmatrix} \cos(\theta) & \text{sen}(\theta) \\ -\text{sen}(\theta) & \cos(\theta) \end{pmatrix}$, logo $\mathbf{M}^T = \begin{pmatrix} \cos(\theta) & -\text{sen}(\theta) \\ \text{sen}(\theta) & \cos(\theta) \end{pmatrix} \Rightarrow \mathbf{M}\,\mathbf{M}^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$.

Observando-se que os vetores coluna da matriz $\mathbf{M}$ são exatamente os componentes dos vetores $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ e $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ no novo sistema de coordenadas, pois $\mathbf{e}_1^{(y)} = \mathbf{M}\,\mathbf{e}_1 = \mathbf{m}^{\langle 1 \rangle} = \begin{pmatrix} \cos(\theta) \\ -\text{sen}(\theta) \end{pmatrix}$ e $\mathbf{e}_2^{(y)} = \mathbf{M}\,\mathbf{e}_2 = \mathbf{m}^{\langle 2 \rangle} = \begin{pmatrix} \text{sen}(\theta) \\ \cos(\theta) \end{pmatrix}$, como mostrado na Figura A.2. ∎

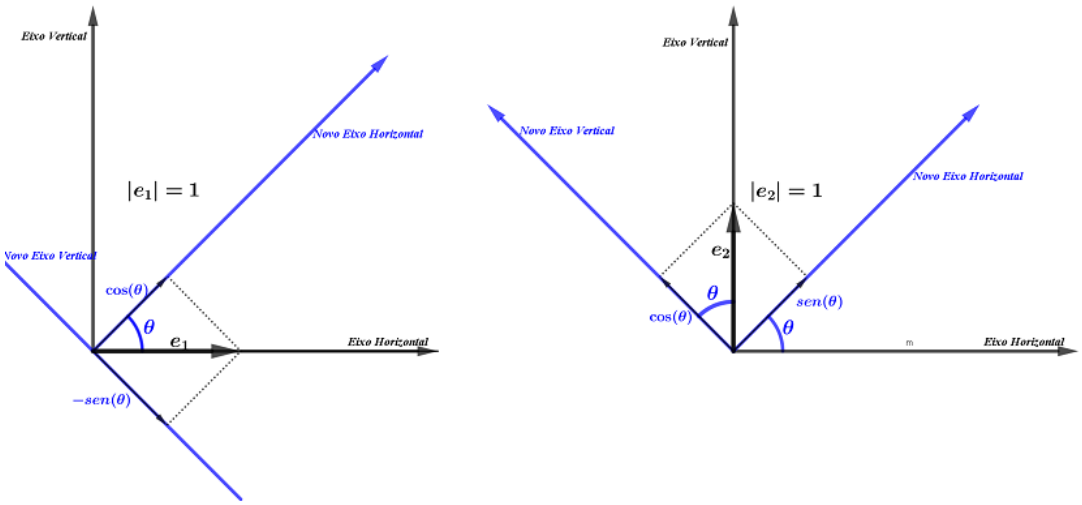


Figura A.2: Coordenadas da base canônica no novo sistema de coordenadas.

Outras propriedades de operações entre matrizes são listadas a seguir.

1. As leis de *associação* e de *comutação* são válidas para as operações de adição/subtração entre matrizes, assim $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$ e $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$.
2. As leis de *associação* e de *distribuição* são válidas para a operação de multiplicação entre matrizes, assim $(\mathbf{A}\,\mathbf{B})\,\mathbf{C}=\mathbf{A}\,(\mathbf{B}\,\mathbf{C})$, $\mathbf{A}\,(\mathbf{B}+\mathbf{C})=\mathbf{A}\,\mathbf{B}+\mathbf{A}\,\mathbf{C}$ e $(\mathbf{A}+\mathbf{B})\,\mathbf{C}=\mathbf{A}\,\mathbf{C}+\mathbf{B}\,\mathbf{C}$.
3. As operações envolvendo matrizes transpostas apresentam as propriedades $(\mathbf{A}+\mathbf{B})^T=\mathbf{A}^T+\mathbf{B}^T$ e $(\mathbf{A}\,\mathbf{B})^T=\mathbf{B}^T\,\mathbf{A}^T$.
4. As operações envolvendo inversas de matrizes apresentam as propriedades $(\mathbf{A}\,\mathbf{B})^{-1}=\mathbf{B}^{-1}\,\mathbf{A}^{-1}$ e $(\mathbf{A}^{-1})^T=(\mathbf{A}^T)^{-1}$.

A.3 Conceito de Posto de Matriz e a Ortogonalização de Gram-Schmidt

Um *menor de ordem p* de uma matriz $\mathbf{A}(n,n)$ é o valor do determinante da matriz obtida eliminando-se $(n-p)$ linhas e $(n-p)$ colunas da matriz $\mathbf{A}$. Se *todos* os menores de ordem $(r+1)$ de uma matriz $\mathbf{A}$ forem nulos e se *ao menos* um menor de ordem r for não nulo, diz-se que a matriz $\mathbf{A}$ é de *posto* (ou *rank*) r. De acordo com esta definição toda matriz quadrada $\mathbf{A}(n,n)$ não singular apresenta o posto igual a n.

Um conjunto de n vetores $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_n$ com n elementos é dito *linearmente independente* se os únicos valores de $c_1,c_2,\cdots,c_n$ tais que $c_1\mathbf{u}_1+c_2\mathbf{u}_2+\cdots+c_n\mathbf{u}_n=\mathbf{0}$ são $c_1=c_2=\cdots=c_n=0$. Neste caso os vetores $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_n$ formam uma *base* do $\Re^n$ e *todo* vetor desse espaço de *dimensão* n (que é o numero máximo de vetores linearmente independentes que pode existir nesse espaço, que também é igual ao número de elementos desses vetores) pode ser expresso como uma combinação linear dos vetores da base, os coeficientes dessa combinação linear são os *componentes* do vetor nessa base. Os componentes de um vetor qualquer do $\Re^n$ apenas confundem-se com seus elementos quando adota-se a *base canônica* do $\Re^n$, que é a base composta pelos vetores unitários $\mathbf{e}_i$ cujo único elemento não nulo é o i-ésimo, isto é, $e_{ij}=\delta_{ij}$. Dessa forma, os vetores coluna ou os vetores linha da matriz identidade $\mathbf{I}$ são os vetores da base canônica do $\Re^n$.

Em uma matriz de posto igual a r todos seus vetores linha (ou coluna) podem ser escritos como uma combinação linear de r vetores linha (ou coluna). Dessa forma, o posto de uma matriz é também o número máximo de vetores linha (ou coluna) linearmente independentes que a matriz contém.

Uma forma de determinar o posto de uma matriz é através do processo de *ortogonalização de Gram*[1]*-Schmidt*[2] aplicado aos vetores linha ou aos vetores coluna da matriz. Este processo pode ser resumido na seguinte forma: sejam $\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_n$ os vetores coluna (ou linha) de $\mathbf{A}$, a partir desses vetores constrói-se uma *base ortogonal* na forma:

$$\mathbf{u}_1=\frac{\mathbf{v}_1}{|\mathbf{v}_1|}$$

$$\mathbf{u}_2=\mathbf{v}_2-(\mathbf{v}_2^T\,\mathbf{u}_1)\,\mathbf{u}_1 \text{ se } |\mathbf{u}_2|>\varepsilon \text{ então } \mathbf{u}_2=\frac{\mathbf{u}_2}{|\mathbf{u}_2|}$$

$$\mathbf{u}_3=\mathbf{v}_3-(\mathbf{v}_3^T\,\mathbf{u}_1)\,\mathbf{u}_1-(\mathbf{v}_3^T\,\mathbf{u}_2)\,\mathbf{u}_2 \text{ se } |\mathbf{u}_3|>\varepsilon \text{ então } \mathbf{u}_3=\frac{\mathbf{u}_3}{|\mathbf{u}_3|}$$

ou, na forma recursiva:

$$\mathbf{u}_j=\mathbf{v}_j-\sum_{k=1}^{j-1}\left[(\mathbf{v}_j^T\,\mathbf{u}_k)\,\mathbf{u}_k\right] \text{ se } |\mathbf{u}_j|>\varepsilon \text{ então } \mathbf{u}_j=\frac{\mathbf{u}_j}{|\mathbf{u}_j|} \text{ para } j=2,\cdots,n \text{ com } \mathbf{u}_1=\frac{\mathbf{v}_1}{|\mathbf{v}_1|},$$

sendo $\varepsilon\approx 0$.

[1] Jorgen Pedersen Gram (1850-1916).
[2] Erhard Schmidt (1876-1959).

Se durante este processo algum vetor $\mathbf{u}_k$ com módulo nulo (ou menor que um valor positivo próximo de zero ε) for encontrado, esse vetor é desconsiderado e o procedimento reiniciado com a renumeração dos vetores subsequentes, ao final do processo o número de vetores $\mathbf{u}_k$ não nulos é igual ao posto da matriz. No final do procedimento um conjunto de r vetores linearmente independentes de dimensão n é encontrado, sendo esses vetores mutuamente ortogonais e de módulos unitários (designados como *ortonormais*). Esse procedimento pode ser também aplicado a matrizes não quadradas.

■ **Exemplo A.2** Calcular através do método de Gram-Schmidt o posto de cada uma das matrizes:

(a) $\mathbf{A} = \begin{pmatrix} 2 & -3 & 7 \\ 4 & 6 & 2 \\ -4 & 0 & 1 \end{pmatrix} \Rightarrow \mathbf{v}_1 = \begin{pmatrix} 2 \\ 4 \\ -4 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix}$ e $\mathbf{v}_3 = \begin{pmatrix} 7 \\ 2 \\ 1 \end{pmatrix}$.

$$\mathbf{u}_1 = \frac{1}{3}\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \Rightarrow \mathbf{v}_2^T\,\mathbf{u}_1 = \mathbf{v}_3^T\,\mathbf{u}_1 = 3$$

$$\mathbf{u}_2 = \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} - \frac{3}{3}\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} -4 \\ 4 \\ 2 \end{pmatrix} \Rightarrow |\mathbf{u}_2| = 6 \Rightarrow \mathbf{u}_2 = \frac{1}{3}\begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} \text{ e } \mathbf{v}_3^T\,\mathbf{u}_2 = -3$$

$$\mathbf{u}_3 = \begin{pmatrix} 7 \\ 2 \\ 1 \end{pmatrix} - \frac{3}{3}\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} + \frac{3}{3}\begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 4 \end{pmatrix} \Rightarrow |\mathbf{u}_3| = 6 \Rightarrow \mathbf{u}_3 = \frac{1}{3}\begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

Como os três vetores $\mathbf{u}_1, \mathbf{u}_2$ e $\mathbf{u}_3$ não são nulos o posto da matriz é igual a 3.

(b) $\mathbf{A} = \begin{pmatrix} -1 & 2 & 3 \\ -2 & 4 & -1 \\ -1 & 2 & -4 \\ -5 & 10 & -6 \end{pmatrix} \Rightarrow \mathbf{v}_1 = -\begin{pmatrix} 1 \\ 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 4 \\ 2 \\ 10 \end{pmatrix}$ e $\mathbf{v}_3 = \begin{pmatrix} 3 \\ -1 \\ -4 \\ -6 \end{pmatrix}$.

Aplicando o método de Gram-Schmidt obtêm-se

$$\mathbf{u}_1 = -\frac{1}{\sqrt{31}}\begin{pmatrix} 1 \\ 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ e } \mathbf{u}_3 = \frac{1}{\sqrt{527}}\begin{pmatrix} 18 \\ 5 \\ -13 \\ -3 \end{pmatrix}.$$

Como há apenas 2 vetores não nulos, o posto desta matriz é igual a 2.

(c) $\mathbf{A} = \begin{pmatrix} -1 & 2 & 3 & 1 \\ -2 & 4 & -1 & -3 \\ -1 & 2 & -4 & -4 \\ -5 & 10 & -6 & -10 \end{pmatrix} \Rightarrow \mathbf{v}_1 = -\begin{pmatrix} 1 \\ 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 4 \\ 2 \\ 10 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 3 \\ -1 \\ -4 \\ -6 \end{pmatrix}$ e $\mathbf{v}_4 = \begin{pmatrix} 1 \\ -3 \\ -4 \\ -10 \end{pmatrix}$.

Aplicando o método de Gram-Schmidt obtêm-se

$$\mathbf{u}_1 = -\frac{1}{\sqrt{31}}\begin{pmatrix} 1 \\ 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{u}_3 = \frac{1}{\sqrt{527}}\begin{pmatrix} 18 \\ 5 \\ -13 \\ -3 \end{pmatrix} \text{ e } \mathbf{u}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Novamente, como há apenas 2 vetores não nulos, o posto desta matriz é igual a 2.

■

A.4 Valores e Vetores Característicos de Matrizes

Dada uma matriz quadrada $\mathbf{A}$ pode-se determinar um escalar λ e um vetor $\mathbf{v}$ tal que a equação $\mathbf{A}\,\mathbf{v} = \lambda\mathbf{v}$ seja satisfeita, o escalar λ é chamado de *valor característico* ou *autovalor* da matriz $\mathbf{A}$ e $\mathbf{v}$ é chamado de *vetor característico* ou *autovetor* de $\mathbf{A}$. A equação de definição do valor e vetor característico pode também ser escrita na forma $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}$, transformando-se assim em um sistema linear e homogêneo de equações que apresenta solução apenas se a matriz for singular, isto

é se a matriz $(\mathbf{A}-\lambda\mathbf{I})$ for singular, o que implica em:

$$\det(\mathbf{A}-\lambda\mathbf{I}) = \det\begin{pmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{pmatrix} = p(\lambda) = 0,$$

sendo $p(\lambda)$ um polinômio de grau *n*ém λ chamado de *polinômio característico* $\mathbf{A}$ cujas n raízes são os *valores característicos* ou *autovalores* da matriz $\mathbf{A}$. Pela expansão do determinante de $(\mathbf{A}-\lambda\mathbf{I})$ verifica-se que o único termo de grau n e $(n-1)$ em λ é o correspondente ao produto da diagonal principal de $(\mathbf{A}-\lambda\mathbf{I})$ isto é, $(a_{11}-\lambda)(a_{22}-\lambda)\cdots(a_{nn}-\lambda)$ sendo todos os demais termos de grau inferior a $(n-1)$. Além disso, como $p(0) = \det(\mathbf{A})$ o termo independente de λ em $p(\lambda)$ é det($\mathbf{A}$), permitindo assim concluir que $p(\lambda) = (-\lambda)^n + (a_{11}+a_{22}+\cdots+a_{nn})(-\lambda)^{n-1} + \cdots + \det(\mathbf{A})$ multiplicando ambos os lados da última equação por $(-1)^n$ obtém-se:

$$p(\lambda) = \lambda^n - (a_{11}+a_{22}+\cdots+a_{nn})\lambda^{n-1} + \cdots + (-1)^n\det(\mathbf{A})$$

(note que apesar de ter-se multiplicado ambos os lados da expressão original por $(-1)^n$, a notação $p(\lambda)$ para designar o polinômio característico foi mantida, pois o polinômio está igualado a zero sendo assim irrelevante seu sinal). Pela expressão de $p(\lambda)$ deduz-se que:

(a) $\lambda_1+\lambda_2+\cdots+\lambda_n = a_{11}+a_{22}+\cdots+a_{nn}$ ou seja $\sum_{i=1}^{n}\lambda_i = \text{tr}(\mathbf{A})$;

(b) $\lambda_1\lambda_2\cdots\lambda_n = \det(\mathbf{A})$ ou seja $\prod_{i=1}^{n}\lambda_i = \det(\mathbf{A})$;

(c) Em decorrência da propriedade (b), se $\mathbf{A}$ for uma matriz singular, em vista de $\det(\mathbf{A}) = 0 \Rightarrow \prod_{i=1}^{n}\lambda_i = 0$, isto é, pelo menos um valor característico de $\mathbf{A}$ é nulo.

Após os valores característicos de $\mathbf{A}$ serem determinados, os vetores característicos são determinados através de $(\mathbf{A}-\lambda_i\mathbf{I})\,\mathbf{v}_i = 0$ para $i = 1, 2, \cdots, n$, porém, sabe-se que para qualquer matriz quadrada $\mathbf{M}$ tem-se $\mathbf{M}\,\text{adj}(\mathbf{M}) = \det(\mathbf{M})\mathbf{I}$, que aplicado a $\mathbf{M} = \mathbf{A}-\lambda_i\mathbf{I}$ em vista de $\det(\mathbf{A}-\lambda_i\mathbf{I}) = 0$, tem-se $(\mathbf{A}-\lambda_i\mathbf{I})\,\text{adj}\,(\mathbf{A}-\lambda_i\mathbf{I}) = \mathbf{0}$ ou ainda $(\mathbf{A}-\lambda_i\mathbf{I})\,[C^{te}\text{adj}\,(\mathbf{A}-\lambda_i\mathbf{I})] = \mathbf{0}$. Desse modo, qualquer vetor coluna não nulo da matriz $\text{adj}\,(\mathbf{A}-\lambda_i\mathbf{I})$ multiplicado por qualquer constante real é um vetor característico de $\mathbf{A}$. Como a matriz adjunta de uma matriz quadrada é obtida transpondo-se a matriz construída substituindo cada elemento por seu cofator, para calcular o vetor característico $\mathbf{v}_i$ basta calcular os cofatores não nulos de uma linha qualquer de $(\mathbf{A}-\lambda_i\mathbf{I})$ multiplicados por uma constante real conveniente.

Pré-multiplicando a equação de definição do valor e vetor característicos pela matriz $\mathbf{A}$ tem-se $\mathbf{A}^2\,\mathbf{v} = \lambda\,(\mathbf{A}\,\mathbf{v})$ e como $\mathbf{A}\,\mathbf{v} = \lambda\mathbf{v}$ tem-se $\mathbf{A}^2\,\mathbf{v} = \lambda^2\mathbf{v}$. Repetindo o procedimento a essa última equação tem-se $\mathbf{A}^3\,\mathbf{v} = \lambda^3\mathbf{v}$, e assim sucessivamente, permitindo escrever: $\mathbf{A}^m\,\mathbf{v} = \lambda^m\mathbf{v}$, para $m = 1, 2, \cdots$, isto é, os valores característicos de $\mathbf{A}^m$ são os valores característicos de $\mathbf{A}$ elevados à mesma potência m e os vetores característicos são iguais aos vetores característicos da matriz $\mathbf{A}$. Se a matriz $\mathbf{A}$ for não singular (não admite o valor característico nulo), então multiplicando ambos os lados de $\mathbf{A}\,\mathbf{v} = \lambda\mathbf{v}$ por $\mathbf{A}^{-1}$, obtém-se $\mathbf{v} = \lambda\left(\mathbf{A}^{-1}\,\mathbf{v}\right) \Rightarrow \mathbf{A}^{-1}\,\mathbf{v} = \dfrac{\mathbf{v}}{\lambda}$, isto é, os valores característicos de $\mathbf{A}^{-1}$ são λ^{-1} e os vetores característicos são iguais aos vetores característicos de $\mathbf{A}$. Dessa forma, pode-se afirmar, se $\mathbf{A}$ for não singular, então os valores característicos de $\mathbf{A}^m$ são os valores característicos de $\mathbf{A}$ elevados à mesma potência m, para $m = 0, \pm1, \pm2, \pm3, \cdots$, e os vetores característicos são iguais aos vetores característicos da matriz $\mathbf{A}$.

Sendo $q(\lambda_i)$ qualquer polinômio escalar de grau n: $q(\lambda_i) = \alpha_n\lambda_i^n + \alpha_{n-1}\lambda_i^{n-1} + \alpha_{n-2}\lambda_i^{n-2} + \cdots + \alpha_1\lambda_i + \alpha_0$ que multiplicado por $\mathbf{v}_i$ resulta em $q(\lambda_i)\mathbf{v}_i = \alpha_n\lambda_i^n\mathbf{v}_i + \alpha_{n-1}\lambda_i^{n-1}\mathbf{v}_i + \alpha_{n-2}\lambda_i^{n-2}\mathbf{v}_i + \cdots + \alpha_1\lambda_i\mathbf{v}_i + \alpha_0\mathbf{v}_i$ mas pela propriedade anterior tem-se $\lambda^m\mathbf{v}_i = \mathbf{A}^m\,\mathbf{v}_i$, resultando em

$q(\lambda_i)\mathbf{v}_i = \left[\alpha_n\mathbf{A}^n + \alpha_{n-1}\mathbf{A}^{n-1} + \alpha_{n-2}\mathbf{A}^{n-2} + \cdots + \alpha_1\mathbf{A} + \alpha_0\mathbf{I}\right]\ \mathbf{v}_i$, ou seja, $q(\mathbf{A})\ \mathbf{v}_i = q(\lambda_i)\mathbf{v}_i$.

Adotando $\alpha_n = (-1)^n$, $\alpha_i = (-1)^n c_i$ para $i = 0, 1, 2, \cdots, (n-1)$, isto é, $q(\lambda_i) = p(\lambda_i)$, o polinômio característico de $\mathbf{A}$, como $p(\lambda_i) = 0$ tem-se:

$$\left[\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I}\right]\ \mathbf{v}_i = \mathbf{0} \text{ para } i = 1, 2, \cdots, n.$$

O Teorema de Cayley[3]-Hamilton[4] pode ser deduzido a partir da análise da equação acima. Considerando que a matriz $\mathbf{A}$ apresenta valores característicos distintos, os correspondentes vetores característicos são linearmente independentes e constituem uma base do espaço vetorial, então qualquer vetor $\mathbf{x}$ desse espaço pode ser expresso por $\mathbf{x} = \sum_{i=1}^{n} b_i\mathbf{v}_i$, então:

$$\left[\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I}\right]\ \sum_{i=1}^{n} b_i\mathbf{v}_i = \mathbf{0}$$

$\left[\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I}\right]\ \mathbf{x} = \mathbf{0}$, como essa expressão é nula para todo $\mathbf{x}$, a matriz $\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I}$ é a matriz nula (todos seus elementos são nulos).

Teorema A.4.1 — Teorema de Cayley-Hamilton. Seja $\mathbf{A}$ uma matriz (n,n) cujo polinômio característico é dado por:

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = (-1)^n\left(\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_2\lambda^2 + c_1\lambda + c_0\right) = 0,$$

então a matriz $\mathbf{A}$ satisfaz seu polinômio característico, ou seja,

$$\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + \cdots + c_2\mathbf{A}^2 + c_1\mathbf{A} + c_0\mathbf{I} = \mathbf{0}.$$

Um procedimento de determinação recursiva dos coeficientes do polinômio característico da $\mathbf{A}$ é desenvolvido baseado na propriedade: $S_m = \sum_{i=1}^{n}\lambda_i^m = \text{tr}(\mathbf{A}^m)$ para $m = 1, 2, \cdots, n$. Considerando a expansão do polinômio característico da $\mathbf{A}$: $p(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_2\lambda^2 + c_1\lambda + c_0$, que pode também ser expresso pelo produto dos monômios $(\lambda - \lambda_i)$ para $i = 1, 2, \cdots, n$ (em que λ_i são os valores característicos de $\mathbf{A}$), $p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$.

Diferenciando ambas expressões de $p(\lambda)$ em relação à λ, obtêm-se:

$p'(\lambda) = n\lambda^{n-1} + (n-1)c_{n-1}\lambda^{n-2} + (n-2)c_{n-2}\lambda^{n-3} + \cdots + 2c_2\lambda + c_1$ e

$p'(\lambda) = (\lambda - \lambda_2)(\lambda - \lambda_3)\cdots(\lambda - \lambda_n) + (\lambda - \lambda_1)(\lambda - \lambda_3)\cdots(\lambda - \lambda_n) + \cdots +$

$+(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_{n-1}) = \sum_{i=1}^{n}\frac{p(\lambda)}{\lambda - \lambda_i}$.

Representando a divisão $\frac{p(\lambda)}{\lambda - \lambda_i}$ por:

$\frac{p(\lambda)}{\lambda - \lambda_i} = \lambda^{n-1} + d_{n-1}^{(i)}\lambda^{n-2} + d_{n-2}^{(i)}\lambda^{n-3} + d_{n-3}^{(i)}\lambda^{n-4} + \cdots + d_2^{(i)}\lambda + d_1^{(i)}$, assim

$p(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_2\lambda^2 + c_1\lambda + c_0 = (\lambda - \lambda_i)(\lambda^{n-1} + d_{n-1}^{(i)}\lambda^{n-2} + d_{n-2}^{(i)}\lambda^{n-3} + d_{n-3}^{(i)}\lambda^{n-4} + \cdots + d_2^{(i)}\lambda + d_1^{(i)}) = \lambda^n + (d_{n-1}^{(i)} - \lambda_i)\lambda^{n-1} + (d_{n-2}^{(i)} - \lambda_i d_{n-1}^{(i)})\lambda^{n-2} + (d_{n-3}^{(i)} - \lambda_i d_{n-2}^{(i)})\lambda^{n-3} + \cdots + (d_1^{(i)} - \lambda_i d_2^{(i)})\lambda - d_1^{(i)}\lambda_i$.

Igualando os termos de mesma potência de λ, resulta:

[3] Arthur Cayley (1821-1895).

[4] William Rowan Hamilton (1805-1865).

$$\begin{cases} d_{n-1}^{(i)} = c_{n-1} + \lambda_i \\ d_{n-2}^{(i)} = c_{n-2} + \lambda_i d_{n-1}^{(i)} \\ d_{n-3}^{(i)} = c_{n-3} + \lambda_i d_{n-2}^{(i)} \\ d_{n-4}^{(i)} = c_{n-4} + \lambda_i d_{n-3}^{(i)} \\ \vdots\vdots\vdots \\ d_2^{(i)} = c_2 + \lambda_i d_3^{(i)} \\ d_1^{(i)} = c_1 + \lambda_i d_2^{(i)} \end{cases} \Rightarrow \begin{cases} d_{n-1}^{(i)} = c_{n-1} + \lambda_i \\ d_{n-2}^{(i)} = c_{n-2} + \lambda_i c_{n-1} + \lambda_i^2 \\ d_{n-3}^{(i)} = c_{n-3} + \lambda_i c_{n-2} + \lambda_i^2 c_{n-1} + \lambda_i^3 \\ d_{n-4}^{(i)} = c_{n-4} + \lambda_i c_{n-3} + \lambda_i^2 c_{n-2} + \lambda_i^3 c_{n-1} + \lambda_i^4 \\ \vdots\vdots\vdots \\ d_2^{(i)} = c_2 + \lambda_i c_3 + \lambda_i^2 c_4 + \lambda_i^3 c_5 + \cdots + \lambda_i^{n-3} c_{n-1} + \lambda_i^{n-2} \\ d_1^{(i)} = c_1 + \lambda_i d_2 + \lambda_i^2 c_3 + \lambda_i^3 c_4 + \cdots + \lambda_i^{n-2} c_{n-1} + \lambda_i^{n-1} \end{cases}$$

e substituindo os valores de $d_n^{(i)}$ em $p'(\lambda) = \sum_{i=1}^{n} \frac{p(\lambda)}{\lambda - \lambda_i}$, tem-se:

$p'(\lambda) = n\lambda^{n-1} + (nc_{n-1} + S_1)\lambda^{n-2} + (c_{n-2} + c_{n-1}S_1 + S_2)\lambda^{n-3} + (nc_{n-3} + c_{n-2}S_1 + c_{n-1}S_2 + S_3)\lambda^{n-4} + \cdots + (nc_2 + c_3S_1 + c_4S_2 + \cdots + c_{n-3}S_{n-5} + c_{n-2}S_{n-4} + c_{n-1}S_{n-3} + S_{n-2})\lambda + (nc_1 + c_2S_1 + c_3S_2 + \cdots + c_{n-3}S_{n-4} + c_{n-2}S_{n-3} + c_{n-1}S_{n-2} + S_{n-1})$, em que $S_m = \sum_{i=1}^{n} \lambda_i^m = \mathrm{tr}(\mathbf{A}^m)$.

Subtraindo da última expressão aquela obtida derivando diretamente a forma original de $p(\lambda)$, ou seja, $p'(\lambda) = n\lambda^{n-1} + (n-1)c_{n-1}\lambda^{n-2} + (n-2)c_{n-2}\lambda^{n-3} + \cdots + ic_i\lambda^{i-1} + \cdots + c_1$, obtêm-se:

$$\begin{cases} c_{n-1} + S_1 = 0 \\ 2c_{n-2} + c_{n-1}S_1 + S_2 = 0 \\ 3c_{n-3} + c_{n-2}S_1 + c_{n-1}S_2 + S_3 = 0 \\ \vdots \\ (n-2)c_2 + c_3S_1 + c_4S_2 + \cdots + c_{n-3}S_{n-5} + c_{n-2}S_{n-4} + c_{n-1}S_{n-3} + S_{n-2} = 0 \\ (n-1)c_1 + c_2S_1 + c_3S_2 + \cdots + c_{n-3}S_{n-4} + c_{n-2}S_{n-3} + c_{n-1}S_{n-2} + S_{n-1} = 0 \end{cases}$$

Além dessas equações tem-se em vista de $p(\lambda_i) = \lambda_i^n + c_{n-1}\lambda_i^{n-1} + c_{n-2}\lambda_i^{n-2} + \cdots + c_2\lambda_i^2 + c_1\lambda_i + c_0 = 0$, que submetida ao somatório de $i = 1$ a n resulta em $nc_0 + c_1S_1 + c_2S_2 \cdots + c_{n-4}S_{n-4} + c_{n-3}S_{n-3} + c_{n-2}S_{n-2} + S_{n-1} = 0$.

Dando origem a um sistema linear triangular de dimensão n cuja solução pode ser expressa na forma recursiva:

$$\begin{cases} c_{n-1} = -S_1 \\ c_{n-2} = -\dfrac{c_{n-1}S_1 + S_2}{2} \\ c_{n-3} = -\dfrac{c_{n-2}S_1 + c_{n-1}S_2 + S_3}{3} \\ \vdots \\ c_2 = -\dfrac{c_3S_1 + c_4S_2 + \cdots + c_{n-3}S_{n-5} + c_{n-2}S_{n-4} + c_{n-1}S_{n-3} + S_{n-2}}{n-2} \\ c_1 = -\dfrac{c_2S_1 + c_3S_2 + \cdots + c_{n-3}S_{n-4} + c_{n-2}S_{n-3} + c_{n-1}S_{n-2} + S_{n-1}}{n-1} \\ c_0 = -\dfrac{c_1S_1 + c_2S_2 \cdots + c_{n-4}S_{n-4} + c_{n-3}S_{n-3} + c_{n-2}S_{n-2} + S_{n-1}}{n} \end{cases}.$$

Este método é chamado de *método de Le Verrier*[5], que determina recursivamente os coeficientes de $p(\lambda)$ a partir do cálculo dos traços das sucessivas potências de $\mathbf{A}$ de 1 a n.

■ **Exemplo A.3** Aplique o método de Le Verrier para determinar o polinômio característico, os valores característicos e os vetores característicos da matriz $\mathbf{A} = \begin{pmatrix} -2,50 & -1,00 & -1,25 \\ -0,50 & -4,00 & -1,75 \\ 1,00 & 2,00 & 0,50 \end{pmatrix}$.

[5] Urbain Jean Joseph Le Verrier (1811–1877).

Assim $\mathbf{A}^2 = \begin{pmatrix} 5,50 & 4,00 & 4,25 \\ 1,50 & 13,00 & 6,75 \\ -3,00 & -8,00 & -4,50 \end{pmatrix}$ e $\mathbf{A}^3 = \begin{pmatrix} -11,50 & -13,00 & -11,75 \\ -3,50 & -40,00 & -21,25 \\ 7,00 & 26,00 & 15,50 \end{pmatrix}$.

Obtêm-se $S_1 = \text{tr}(\mathbf{A}) = -6$, $S_2 = \text{tr}(\mathbf{A})^2 = 14$ e $S_3 = \text{tr}(\mathbf{A})^3 = -36$.

$$\begin{cases} c_2 = -S_1 = 6 \\ c_1 = -\dfrac{c_2 S_1 + S_2}{2} = -\dfrac{6(-6)+14}{2} = 11 \\ c_0 = -\dfrac{c_1 S_1 + c_2 S_2 + S_3}{3} = -\dfrac{11(-6)+6(14)-36}{3} = 6 \end{cases}.$$

Verificando-se que $\mathbf{A}^3 + 6\mathbf{A}^3 + 11\mathbf{A} + 6\mathbf{I} = \mathbf{0}$, provando que os valores dos coeficientes estão corretos, pois satisfazem ao estabelecido no Teorema de Cayley-Hamilton. O polinômio característico de $\mathbf{A}$ é então $p(\lambda) = \lambda^3 + 6\lambda^2 + 11\lambda + 6$, cujas raízes (os valores característicos de $\mathbf{A}$) são: $\lambda_1 = -1$, $\lambda_2 = -2$ e $\lambda_3 = -3$. Os vetores característicos de $\mathbf{A}$ são determinados de acordo com:

1. Primeiro vetor característico: correspondendo a $\lambda_1 = -1$

   $\mathbf{A} - \lambda_1\mathbf{I} = \mathbf{A} + \mathbf{I} = \begin{pmatrix} -1,50 & -1,00 & -1,25 \\ -0,50 & -3,00 & -1,75 \\ 1,00 & 2,00 & 1,50 \end{pmatrix}$ cujos cofatores da primeira linha são -1, -1 e 2.

   Multiplicando estes valores por -1 obtém-se $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$.

2. Segundo vetor característico: correspondendo a $\lambda_1 = -2$

   $\mathbf{A} - \lambda_2\mathbf{I} = \mathbf{A} + 2\mathbf{I} = \begin{pmatrix} -0,50 & -1,00 & -1,25 \\ -0,50 & -2,00 & -1,75 \\ 1,00 & 2,00 & 2,50 \end{pmatrix}$ cujos cofatores da primeira linha são $-1,5$; $-0,5$ e 1. Multiplicando estes valores por -2 obtém-se $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 1 \\ -2 \end{pmatrix}$.

3. Terceiro vetor característico: correspondendo a $\lambda_1 = -3$

   $\mathbf{A} - \lambda_3\mathbf{I} = \mathbf{A} + 3\mathbf{I} = \begin{pmatrix} 0,50 & -1,00 & -1,25 \\ -0,50 & -1,00 & -1,75 \\ 1,00 & 2,00 & 3,50 \end{pmatrix}$ cujos cofatores da primeira linha são 1, 3 e -2, obtém-se $\mathbf{v}_3 = \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$.

■

Outra propriedade importante relativa a valores e vetores característicos de matrizes diz respeito aos valores e vetores característicos de $\mathbf{A}^T$. Assim, seja μ_i um valor característico de $\mathbf{A}^T \Rightarrow \mathbf{A}^T\,\mathbf{u}_i = \mu_i\mathbf{u}_i$, sendo $\mathbf{u}_i$ o vetor característico correspondente, então μ_i são as raízes de $\det(\mathbf{A}^T - \mu\mathbf{I}) = 0$, mas $\det(\mathbf{A}^T - \mu\mathbf{I})$=$\det[(\mathbf{A}^T - \mu\mathbf{I})]^T$=$\det(\mathbf{A} - \mu\mathbf{I})$ que é idêntico ao polinômio característico de $\mathbf{A}$, demonstrando que os valores característicos de $\mathbf{A}^T$ são iguais aos valores característicos de $\mathbf{A}$. Entretanto, o mesmo não ocorre com os vetores característicos; se $\mathbf{v}_i$, para $i = 1, 2, \cdots, n$, forem os vetores característicos de $\mathbf{A}$ e $\mathbf{u}_j$, para $j = 1, 2, \cdots, n$, forem os vetores característicos de $\mathbf{A}^T$ tem-se:

$$\begin{cases} \mathbf{A}\,\mathbf{v}_i = \lambda_i\mathbf{v}_i \\ \mathbf{A}^T\,\mathbf{u}_j = \lambda_j\mathbf{u}_j \Rightarrow \mathbf{u}_j^T\,\mathbf{A} = \lambda_j\mathbf{u}_j^T \text{ para } j \neq i \end{cases},$$

assim, $\mathbf{u}_j^T\,\mathbf{A}\,\mathbf{v}_i = \lambda_j\mathbf{u}_j^T\,\mathbf{v}_i = \lambda_i\mathbf{u}_j^T\,\mathbf{v}_i \Rightarrow (\lambda_j - \lambda_i)\mathbf{u}_j^T\,\mathbf{v}_i = 0$ como $\lambda_j \neq \lambda_i \Rightarrow \mathbf{u}_j^T\,\mathbf{v}_i = 0$, isto é, os vetores característicos de $\mathbf{A}^T$ e de $\mathbf{A}$, correspondendo a valores característicos distintos, são ortogonais entre si.

Consequentemente, matrizes simétricas $(\mathbf{A} = \mathbf{A}^T)$ apresentam vetores característicos ortogonais entre si, correspondentes a valores característicos distintos. Além disto, os valores característicos de matrizes simétricas são todos reais, esta propriedade pode ser demonstrada considerando hipótese contrária. Isto é, sendo $\lambda_k = \alpha + \beta i$ um valor característico de uma matriz simétrica $\mathbf{A}$, então, como a matriz e os coeficientes do polinômio característico são todos reais, $\bar{\lambda}_k = \alpha - \beta i$ (seu conjugado) também será valor característico de $\mathbf{A}$. A mesma propriedade ocorreria com os correspondentes vetores característicos, $\mathbf{v}_k = \mathbf{a} + \mathbf{b}i$ e $\bar{\mathbf{v}}_k = \mathbf{a} - \mathbf{b}i$, no entanto, $(\bar{\lambda}_k - \lambda_k)\bar{\mathbf{v}}_k^T\ \mathbf{v}_k = 2\beta\bar{\mathbf{v}}_k^T\ \mathbf{v}_k = 0$ sendo $\bar{\mathbf{v}}_k^T\ \mathbf{v}_k = (\mathbf{a} + \mathbf{b}i)^T\ (\mathbf{a} + \mathbf{b}i) = ||\mathbf{a}||^2 + ||\mathbf{b}||^2$ logo $2\beta\bar{\mathbf{v}}_k^T\ \mathbf{v}_k = 2\beta\left(||\mathbf{a}||^2 + ||\mathbf{b}||^2\right) = 0$, como $\left(||\mathbf{a}||^2 + ||\mathbf{b}||^2\right) > 0 \Rightarrow \beta = 0$, contradizendo a hipótese da matriz admitir um valor característico complexo.

A.5 Valores e Vetores Singulares

Dada uma matriz $\mathbf{A} \in \Re^{m\times n}$ e lembrando dos conceitos básicos que a imagem ou range($\mathbf{A}$) = $\{\mathbf{A}\,\mathbf{x} \mid \mathbf{x} \in \Re^n\}$ é o espaço gerado pelos vetores colunas de $\mathbf{A}$ e range($\mathbf{A}^T$) é o espaço gerado pelos vetores linhas de $\mathbf{A}$, a decomposição em valores e vetores singulares (*Singular Value Decomposition*, SVD) é capaz de obter simultaneamente as bases ortonormais desses subespaços.

Qualquer matriz $\mathbf{A} \in \Re^{m\times n}$ pode ser decomposta na forma:

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^T,$$

em que $\mathbf{U} \in \Re^{m\times m}$ é uma matriz ortonormal cujas colunas são os vetores característicos de $\mathbf{A}\,\mathbf{A}^T \in \Re^{m\times m}$, $\mathbf{V} \in \Re^{n\times n}$ é uma matriz ortonormal cujas colunas são os vetores característicos de $\mathbf{A}^T\,\mathbf{A} \in \Re^{n\times n}$ e $\Sigma \in \Re^{m\times n}$ é uma *matriz diagonal retangular* contendo a raiz quadrada dos valores característicos de $\mathbf{A}\,\mathbf{A}^T$ (que são equivalentes aos valores característicos de $\mathbf{A}^T\,\mathbf{A}$), arranjados em ordem decrescente. As últimas linhas ou colunas excedentes na matriz *Sigma* em relação à matriz diagonal quadrada de dimensão $\max(n,m)$ contêm somente elementos nulos. Os vetores característicos de $\mathbf{A}\,\mathbf{A}^T$ e $\mathbf{A}^T\,\mathbf{A}$ estão arranjados nas colunas de $\mathbf{U}$ e $\mathbf{V}$, respectivamente, na ordem de seus valores característicos na matriz Σ. Os elementos, σ_i, da diagonal de Σ são denominados de *valores singulares* de $\mathbf{A}$, sendo todos não negativos. Além disso, o número de valores singulares positivos é igual ao posto($\mathbf{A}$). Os vetores colunas de $\mathbf{U}$ são denominados de *vetores singulares à esquerda* de $\mathbf{A}$ e os vetores colunas de $\mathbf{V}$ são denominados de *vetores singulares à direita* de $\mathbf{A}$, e as relações entre esses vetores são:

$$\mathbf{A}\,\mathbf{v}_i = \sigma_i\,\mathbf{u}_i \text{ e } \mathbf{A}^T\,\mathbf{u}_i = \sigma_i\,\mathbf{v}_i.$$

A decomposição SVD revela várias propriedades intrínsecas da matriz $\mathbf{A}$ e é numericamente estável para os cálculos. Algumas propriedades são listadas abaixo para uma matriz com r valores singulares positivos:

(a) posto($\mathbf{A}$) = r;
(b) null($\mathbf{A}$) = span($\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \cdots, \mathbf{v}_n$);
(c) range($\mathbf{A}$) = span($\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r$);
(d) range($\mathbf{A}^T$) = span($\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r$);
(e) $\mathbf{A} = \sum_{i=1}^{r} \sigma_i\,\mathbf{u}_i\,\mathbf{v}_i^T$;
(f) $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$ (norma de Frobenius);
(g) $\|\mathbf{A}\|_2 = \sigma_1$.

em que null($\mathbf{A}$) = $\{\mathbf{x} \in \Re^n \mid \mathbf{A}\,\mathbf{x} = \mathbf{0}\} \subseteq \Re^n$ é o espaço nulo da matriz $\mathbf{A}$ e span$(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r) = \sum_{i=1}^{r} \alpha_i \mathbf{u}_i$ é o subespaço gerado por todas as combinações lineares dos vetores do conjunto gerador.

A matriz: $\mathbf{A}^\dagger = \mathbf{V}\,\Sigma^\dagger\,\mathbf{U}^T = (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T$, é chamada de *pseudo-inversa* de $\mathbf{A}$, em que os elementos da diagonal de $\Sigma^\dagger$ consistem no recíproco dos valores singulares positivos de Σ, na mesma ordem. A pseudo-inversa tem a propriedade $\mathbf{A}\,\mathbf{A}^\dagger\,\mathbf{A} = \mathbf{A}$ ou $\mathbf{A}^\dagger\,\mathbf{A}\,\mathbf{A}^\dagger = \mathbf{A}^\dagger$.

A solução do problema de valores singulares para o sistema linear, $\mathbf{y} = \mathbf{A}\,\mathbf{x}$, corresponde resolver o seguinte problema de otimização:

$$\begin{array}{l} \max\limits_{\mathbf{x}} ||\mathbf{y}||_2^2 \\ \text{sujeito a: } ||\mathbf{x}||_2^2 = 1, \end{array}$$

em que $||\mathbf{y}||_2^2 = \mathbf{y}^T\,\mathbf{y}$. Usando o conceito dos multiplicadores de Lagrange, o problema acima pode ser reescrito como:

$$\max_{\mathbf{x}} [S(\mathbf{x}) = \mathbf{x}^T\,\mathbf{A}^T\,\mathbf{A}\,\mathbf{x} - \lambda\,(\mathbf{x}^T\,\mathbf{x} - 1)]$$

cuja primeira condição de otimalidade é $\nabla S(\mathbf{x}) = 2\,\mathbf{x}^T\,\mathbf{A}\,\mathbf{x} - 2\,\lambda\,\mathbf{x} = 0$, ou seja, a solução é equivalente ao problema de valor característico $\mathbf{A}^T\,\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}$, cujos $\mathbf{x}$ ótimos locais correspondem aos vetores característicos de $\mathbf{A}^T\,\mathbf{A}$ ou os vetores singulares de $\mathbf{A}$ (também chamados de *componentes principais* de variação, pois indicam as direções de máxima variação de $\mathbf{y}$ em função das variações em $\mathbf{x}$ com mesma energia, isto é, $||\mathbf{x}||_2 = 1$ e os respectivos multiplicadores de Lagrange são os valores característicos de $\mathbf{A}^T\,\mathbf{A}$ ou o quadrado dos valores singulares de $\mathbf{A}$.

Observe que para uma matriz de posto$(\mathbf{A}) = r$, $\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i\,\mathbf{u}_i\,\mathbf{v}_i^T$ e, portanto, $\mathbf{y} = \mathbf{A}\,\mathbf{x} = \mathbf{U}\,\Sigma\,\mathbf{V}^T\,\mathbf{x} = \sum_{i=1}^{r} \sigma_i\,\mathbf{u}_i\,\mathbf{v}_i^T\,\mathbf{x}$, indicando que a projeção do vetor $\mathbf{x}$ na direção do vetor $\mathbf{v}_i$ (ou seja, $\mathbf{v}_i^T\,\mathbf{x}$) é amplificada por σ_i na direção $\mathbf{u}_i$ do vetor $\mathbf{y}$, sendo $i = 1$ a direção de maior amplificação e $i = r$ a direção de menor amplificação. Dependendo do valor de σ_i, uma pequena mudança em $\mathbf{x}$ pode causar uma grande mudança em $\mathbf{y}$, mas isto vai depender do ângulo entre os vetores $\mathbf{x}$ e $\mathbf{v}_i$.

■ **Exemplo A.4** Usando vetores unitários $\mathbf{u}$, $||\mathbf{u}|| = 1$, e suas transformações lineares $\mathbf{A}\,\mathbf{u}$ para uma matriz de dimensão 2×2 é possível localizar as direções características da matriz na Figura A.3a.

$$A = \begin{pmatrix} 1/4 & 3/4 \\ 1 & 1/2 \end{pmatrix}$$

em que os valores característicos, λ, e os vetores característicos, $\mathbf{u}$, resultantes das soluções não triviais do sistema de equações lineares $\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u}$, são: $\lambda_1 = 5/4$ e $\mathbf{u}_1 = \frac{1}{5}\begin{pmatrix} 3 \\ 4 \end{pmatrix}$; $\lambda_2 = -1/2$ e $\mathbf{u}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, são visualizados quando os dois vetores ($\mathbf{u}$ e $\mathbf{A}\,\mathbf{u}$) estão na mesma direção. Mostrando que o operador $\mathbf{A}$, na direção de $\mathbf{u}$, corresponde a uma redução ou ampliação por um fator λ. Quando os sentidos dos dois vetores são opostos tem-se um valor característico negativo.

Pode-se observar que os dois vetores característicos não são os eixos maior e menor da elipse formada pelas transformações lineares. Seriam para o caso particular de matrizes simétricas. Matrizes 2×2 com um par de valores característicos complexos não possuem vetores característicos reais.

Usando dois vetores unitários, $\mathbf{v}_1$ e $\mathbf{v}_2$, perpendiculares e suas correspondentes transformações lineares, $\mathbf{A}\,\mathbf{v}_1$ e $\mathbf{A}\,\mathbf{v}_2$, pode-se localizar os valores e vetores singulares, resultantes das soluções não triviais dos sistemas de equações lineares $\mathbf{A}^T\,\mathbf{A}\,\mathbf{v} = \sigma^2\,\mathbf{v}$ e $\mathbf{A}\,\mathbf{A}^T\,\mathbf{u} = \sigma^2\,\mathbf{u}$, que são:

$\sigma_1 = 1{,}2792$, $\mathbf{v}_1 = (0{,}7678 \quad 0{,}6407)^T$ e $\mathbf{A}\,\mathbf{v}_1 = \sigma_1\,\mathbf{u}_1 = (0{,}6725 \quad 1{,}0881)^T$

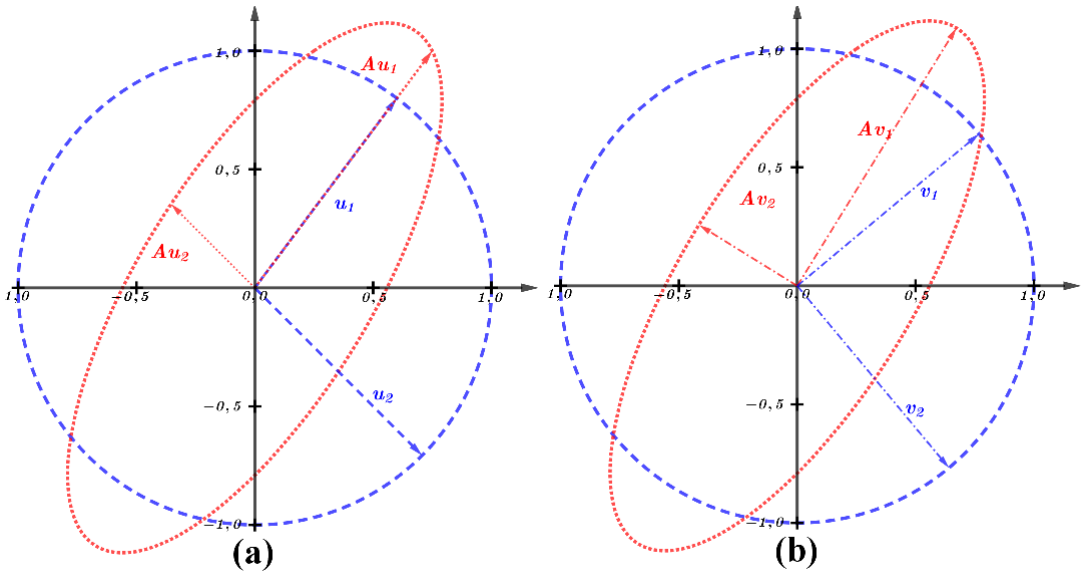


Figura A.3: Direções características de uma matriz.

$\sigma_2 = 0,4886$, $\mathbf{v}_2 = (0,6407 \quad -0,7678)^T$ e $\mathbf{A}\,\mathbf{v}_2 = \sigma_2\,\mathbf{u}_2 = (-0,4156 \quad 0,2569)^T$,
em que σ_1 e σ_2 são os valores singulares, $(\mathbf{v}_1,\ \mathbf{v}_2)$ e $(\mathbf{u}_1, \mathbf{u}_2)$ são as matrizes dos vetores singulares à direita e à esquerda, respectivamente. Esses vetores surgem no momento em que as transformações são perpendiculares entre si, conforme mostra a Figura A.3b. Observa-se que isto acontece quando os vetores das transformações são os eixos maior e menor da elipse, mostrando, por exemplo, as direções de máxima e mínima amplificação de sinais, respectivamente. Para o caso particular de uma matriz quadrada, simétrica e positiva definida, as decomposições em valores característicos e em valores singulares são equivalentes. ∎

Para determinar se o sistema linear, $\mathbf{y} = \mathbf{A}\,\mathbf{x}$, está bem escalonado, faz-se uso do número de condicionamento da matriz $\mathbf{A}$, que na norma 2 é dado por:

$$\gamma(\mathbf{A}) = \frac{\sigma_{max}(\mathbf{A})}{\sigma_{min}(\mathbf{A})}$$

em que $\sigma_{max}(\mathbf{A})$ é o maior valor singular de $\mathbf{A}$ e $\sigma_{min}(\mathbf{A})$ é o menor valor singular não nulo de $\mathbf{A}$.

A melhor maneira de escalonar um sistema é atacando a origem do problema, ou seja um apropriado adimensionamento das variáveis dependentes e das equações do problema. Uma maneira numérica de determinar quais as variáveis devem ser reescaladas é através do cálculo do condicionamento mínimo, isto é, determinar as matrizes que pré- e pós-multiplicadas pela matriz $\mathbf{A}$ resultam em um γ mínimo (γ^*), isto é:

$$\gamma^*(\mathbf{A}) = \min_{L,R} \gamma(\mathbf{L}\,\mathbf{A}\,\mathbf{R}).$$

Considerando $\mathbf{L}$ e $\mathbf{R}$ matrizes diagonais, então tem-se como resultado do problema de otimização acima quais as saídas e entradas devem ser reescaladas, respectivamente, pois:

$$\mathbf{y}_e = \mathbf{L}\,\mathbf{y} \quad \text{e} \quad \mathbf{x}_e = \mathbf{R}^{-1}\,\mathbf{x}.$$

A.6 Formas Canônicas de Matrizes

A utilização dos conceitos de matrizes e vetores é plenamente justificada na representação de sistemas algébricos lineares $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, nos quais tanto os elementos do vetor das incógnitas $\mathbf{x}$

quanto os elementos do vetor das constantes $\mathbf{b}$ são seus componentes na base canônica de $\Re^n$. Os componentes dos vetores $\mathbf{x}$ e $\mathbf{b}$ em uma nova base de $\Re^n$, constituída por n vetores linearmente independentes $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$, seriam determinados através de $\mathbf{x} = \sum_{i=1}^{n} y_i \mathbf{p}_i$ e $\mathbf{b} = \sum_{i=1}^{n} c_i \mathbf{p}_i$, ou, em notação matricial, $\mathbf{x} = \mathbf{P}\mathbf{y}$ e $\mathbf{b} = \mathbf{P}\,\mathbf{c}$, em que $\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \cdots & \mathbf{p}_n \end{pmatrix}$.

O sistema original transforma-se em $\mathbf{A}\,\mathbf{P}\,\mathbf{y} = \mathbf{P}\,\mathbf{c} \Rightarrow (\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})\,\mathbf{y} = \mathbf{c}$, definindo $\mathbf{B} = (\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})$, permitindo interpretar $\mathbf{B}\,\mathbf{y} = \mathbf{c}$ como sendo o sistema original representado na *nova base* $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$.

Um propriedade importante da matriz $\mathbf{B} = (\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})$, que é *similar* (ou *semelhante*) à matriz $\mathbf{A}$, é a *invariância* dos valores característicos de $\mathbf{A}$ na nova base, pois

$\det(\mathbf{B} - \lambda\mathbf{I}) = \det(\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P} - \lambda\mathbf{I}) = \det\left[(\mathbf{P}^{-1}\,(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{P}\right] =$
$= \det(\mathbf{P}^{-1})\det(\mathbf{A} - \lambda\mathbf{I})\det(\mathbf{P}) = \det(\mathbf{P}^{-1})\det(\mathbf{P})\det(\mathbf{A} - \lambda\mathbf{I}) = \det(\mathbf{A} - \lambda\mathbf{I}) = p(\lambda)$. Portanto, o polinômio característico da matriz $\mathbf{B}$ é o mesmo polinômio característico da matriz $\mathbf{A}$, o mesmo ocorrendo em relação aos valores característicos. Entretanto, os vetores característicos de $\mathbf{A}$ e $\mathbf{B}$ não são os mesmos. Sendo $\mathbf{v}$ os vetores característicos de $\mathbf{A}$, tem-se $\mathbf{A}\,\mathbf{v} = \lambda\mathbf{v}$, cujos componentes na nova base são $\mathbf{v} = \sum_{i=1}^{n} u_i \mathbf{p}_i = \mathbf{P}\,\mathbf{u}$, assim $\mathbf{A}\,\mathbf{P}\,\mathbf{u} = \lambda(\mathbf{P}\,\mathbf{u}) \Rightarrow (\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})\,\mathbf{u} = \mathbf{B}\,\mathbf{u} = \lambda\mathbf{u}$, isto é, os vetores característicos da matriz $\mathbf{B}$ nada mais são que a representação dos vetores característicos da matriz $\mathbf{A}$ na nova base.

Outra propriedade da mudança de base de matrizes diz respeito à potenciação, assim, considerando a expressão $\mathbf{Q}^{(k)} = (\mathbf{P}^{-1}\,\mathbf{A}^{\mathbf{k}}\,\mathbf{P})$, que pré-multiplicada por $(\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})$ resulta em $(\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P})\,(\mathbf{P}^{-1}\,\mathbf{A}^{\mathbf{k}}\,\mathbf{P}) = \mathbf{P}^{-1}\,\mathbf{A}^{\mathbf{k+1}}\,\mathbf{P} = \mathbf{Q}^{(k+1)}$. Identificando $\mathbf{Q}^{(1)} = (\mathbf{P}^{-1}\,\mathbf{A}\,\mathbf{P}) = \mathbf{B}$, concluí-se que $\mathbf{B}^m = \mathbf{P}^{-1}\,\mathbf{A}^m\,\mathbf{P}$ e $\mathbf{A}^m = \mathbf{P}\,\mathbf{B}^m\,\mathbf{P}^{-1}$, para $m = 1, 2, \cdots$, e, se $\mathbf{A}$ for não singular, inclui também os valores inteiros negativos.

Esta mesma propriedade pode ser aplicada a funções polinomiais de $\mathbf{A}$ do tipo

$$p_m(\mathbf{A}) = \mathbf{A}^m + a_{m-1}\mathbf{A}^{m-1} + a_{m-2}\mathbf{A}^{m-2} + \cdots + a_1\mathbf{A} + a_0\mathbf{I},$$

implicando em $p_m(\mathbf{B}) = \mathbf{P}^{-1}\,p_m(\mathbf{A})\,\mathbf{P}$ e $p_m(\mathbf{A}) = \mathbf{P}\,p_m(\mathbf{B})\,\mathbf{P}^{-1}$

Se a matriz $\mathbf{A}$ apresentar n valores característicos distintos então a base constituída pelos vetores característicos de $\mathbf{A}$, $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$, compondo a matriz $\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \cdots & \mathbf{v}_n \end{pmatrix}$ em vista de $\mathbf{A}\,\mathbf{v}_i = \lambda_i\mathbf{v}_i$ e $\mathbf{A}\,\mathbf{V} = \begin{pmatrix} \mathbf{A}\,\mathbf{v}_1 & \mathbf{A}\,\mathbf{v}_2 & \mathbf{A}\,\mathbf{v}_3 & \cdots & \mathbf{A}\,\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \lambda_1\mathbf{v}_1 & \lambda_2\mathbf{v}_2 & \lambda_3\mathbf{v}_3 & \cdots & \lambda_n\mathbf{v}_n \end{pmatrix} = \mathbf{V}\,\mathrm{diag}\begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_n \end{pmatrix}$, então $\mathbf{V}^{-1}\mathbf{A}\,\mathbf{V} = \mathrm{diag}\begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_n \end{pmatrix}$.

Desse modo, a matriz $\mathbf{A}$ representada na base composta por seus vetores característicos, assume sua forma mais simples que é a matriz diagonal composta por seus valores característicos (caso forem todos distintos):

$$\mathbf{V}^{-1}\mathbf{A}\,\mathbf{V} = \mathbf{D} = \mathrm{diag}\begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_n \end{pmatrix} \quad \text{e} \quad \mathbf{A} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^{-1}.$$

Nesse caso, a matriz diagonal é a *forma canônica* da matriz $\mathbf{A}$ e o procedimento é chamado de *diagonalização*.

Se a matriz $\mathbf{A}$ apresentar valores característicos múltiplos e se ao(s) valor(es) característico(s) múltiplo(s) associar(em)-se apenas um vetor característico, a *forma canônica* da matriz não é mais uma matriz diagonal mas sim a *Forma Canônica de Jordan*. Para ilustrar o procedimento de determinação da nova base que transforma a matriz $\mathbf{A}$ na correspondente forma canônica, o primeiro valor característico λ_1 é considerado como de multiplicidade m e os $(n-m)$ restantes distintos entre si e diferentes de λ_1. Ao valor característico λ_1 associa-se apenas um vetor característico $\mathbf{v}_1$ tal que $\mathbf{A}\,\mathbf{v}_1 = \lambda_1\mathbf{v}_1$, o posto da matriz $(\mathbf{A} - \lambda_1\mathbf{I})$ é $(n-1)$, e aos demais $(n-m)$ restantes valores característicos associam-se os vetores característicos $\mathbf{v}_k$ que satisfazem a $\mathbf{A}\,\mathbf{v}_k = \lambda_k\mathbf{v}_k$ para $k = (m+1), (m+2), \cdots, n$. Os vetores característicos $\mathbf{v}_1, \mathbf{v}_{m+1}, \mathbf{v}_{m+2}, \cdots, \mathbf{v}_n$, constituem a primeira coluna e as colunas $(m+1), (m+2), \cdots, n$ da matriz $\mathbf{V}$. Para determinar as demais colunas desta matriz (colunas: 2 , 3, ..., m) assim procede-se:

$\mathbf{A}\,\mathbf{v}_j = \lambda_1\mathbf{v}_j + \mathbf{v}_{j-1}$ para $j = 2, 3, \cdots, m$.

Deste modo tem-se $\mathbf{A}\,\mathbf{V} = \begin{pmatrix}\mathbf{A}\,\mathbf{v}_1 & \mathbf{A}\,\mathbf{v}_2 & \mathbf{A}\,\mathbf{v}_3 & \cdots & \mathbf{A}\,\mathbf{v}_n\end{pmatrix}$

$\mathbf{A}\,\mathbf{V} = \begin{pmatrix}\lambda_1\mathbf{v}_1 & \lambda_1\mathbf{v}_2+\mathbf{v}_1 & \lambda_1\mathbf{v}_3+\mathbf{v}_2 & \cdots & \lambda_1\mathbf{v}_m+\mathbf{v}_{m-1} & \lambda_{m+1}\mathbf{v}_{m+1} & \cdots & \lambda_n\mathbf{v}_n\end{pmatrix}$.

Identificando $\begin{pmatrix}\lambda_1\mathbf{v}_1 & \lambda_1\mathbf{v}_2+\mathbf{v}_1 & \lambda_1\mathbf{v}_3+\mathbf{v}_2 & \cdots & \lambda_1\mathbf{v}_m+\mathbf{v}_{m-1} & \lambda_{m+1}\mathbf{v}_{m+1} & \cdots & \lambda_n\mathbf{v}_n\end{pmatrix} =$

$$= \mathbf{V}\begin{pmatrix}\lambda_1 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0\\ 0 & \lambda_1 & 1 & \cdots & 0 & 0 & \cdots & 0\\ 0 & 0 & \lambda_1 & \cdots & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & \lambda_1 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0 & \lambda_{m+1} & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \lambda_n\end{pmatrix} = \mathbf{V}\,\mathbf{J},$$

sendo $\mathbf{J}$ a forma canônica de Jordan da matriz $\mathbf{A}$.

Se ao valor característico de multiplicidade m associar-se mais de um vetor característico, o que ocorre quando o posto da matriz $(\mathbf{A}-\lambda^*\mathbf{I})$ for igual a $(n-k)$ com $1 < k \leq n$, neste caso, ao valor característico λ^* associam-se k vetores característicos linearmente independentes. Para ilustrar o procedimento de determinação da forma canônica da matriz $\mathbf{A}$ neste caso, o primeiro valor característico λ_1 é considerado como de multiplicidade m e os $(n-m)$ restantes distintos entre si e diferentes de λ_1. Ao valor característico λ_1 associam-se k vetores característicos distintos $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$, tal que $\mathbf{A}\,\mathbf{v} = \lambda_1\mathbf{v}$, que apresenta k soluções, o posto da matriz $(\mathbf{A}-\lambda_1\mathbf{I})$ é $(n-k)$, e aos demais $(n-m)$ restantes valores característicos associam-se os vetores característicos $\mathbf{v}_j$ que satisfazem a $\mathbf{A}\,\mathbf{v}_j = \lambda_j\mathbf{v}_j$ para $j = (m+1), (m+2), \cdots, n$. Os vetores característicos $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k, \mathbf{v}_{m+1}, \mathbf{v}_{m+2}, \cdots, \mathbf{v}_n$, constituem as k primeiras colunas e as colunas $(m+1), (m+2), \cdots, n$ da matriz $\mathbf{V}$, para determinar as demais colunas desta matriz (colunas $k+1, k = 2, \cdots m$) assim procede-se:

$\mathbf{A}\,\mathbf{v}_j = \lambda_1\mathbf{v}_j + \mathbf{v}_{j-1}$ para $j = k+1, k+2, \cdots, m$

com $\mathbf{v}_k$ o k-ésimo vetor característico correspondente a λ_1.

Nesse caso, tem-se $\mathbf{A}\,\mathbf{V} = \begin{pmatrix}\mathbf{A}\,\mathbf{v}_1 & \mathbf{A}\,\mathbf{v}_2 & \mathbf{A}\,\mathbf{v}_3 & \cdots & \mathbf{A}\,\mathbf{v}_n\end{pmatrix}$

$\mathbf{A}\,\mathbf{V} = \begin{pmatrix}\lambda_1\mathbf{v}_1 & \lambda_1\mathbf{v}_2 & \cdots & \lambda_1\mathbf{v}_k & \lambda_1\mathbf{v}_{k+1}+\mathbf{v}_k & \cdots & \lambda_1\mathbf{v}_m+\mathbf{v}_{m-1} & \lambda_{m+1}\mathbf{v}_{m+1} & \cdots & \lambda_n\mathbf{v}_n\end{pmatrix}$.

Nessa situação, a forma canônica de Jordan é diferente da apresentada anteriormente, na qual o valor unitário só aparece sobre o elemento da diagonal após a coluna k.

■ **Exemplo A.5** Para ilustrar as diferentes formas canônicas de Jordan consideram-se as seguintes matrizes

(a) $\mathbf{A} = \begin{pmatrix}2 & 3 & 4\\ 0 & 2 & 3\\ 0 & 0 & 2\end{pmatrix} \Rightarrow p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) = \det\begin{pmatrix}2-\lambda & 3 & 4\\ 0 & 2-\lambda & 3\\ 0 & 0 & 2-\lambda\end{pmatrix} = (2-\lambda)^3 \Rightarrow \lambda_1 =$

$\lambda_2 = \lambda_3 = 2$

$\mathbf{A}-2\mathbf{I} = \begin{pmatrix}0 & 3 & 4\\ 0 & 0 & 3\\ 0 & 0 & 0\end{pmatrix}$, posto$(\mathbf{A}-2\mathbf{I}) = 2$ então só há um vetor característico associado a λ_1 que é determinado por:

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_1 = \begin{pmatrix}3v_{12}+4v_{13}\\ 3v_{13}\\ 0\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$.

Então $v_{12} = v_{13} = 0$ e $v_{11} \neq 0 \Rightarrow \mathbf{v}_1 = \begin{pmatrix}1\\ 0\\ 0\end{pmatrix}$.

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_2 = \begin{pmatrix} 3v_{22}+4v_{23} \\ 3v_{23} \\ 0 \end{pmatrix} = \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, então $v_{23}=0$, $v_{22}=\dfrac{1}{3}$ e v_{21} qualquer $\Rightarrow \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1/3 \\ 0 \end{pmatrix}$.

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_3 = \begin{pmatrix} 3v_{32}+4v_{33} \\ 3v_{33} \\ 0 \end{pmatrix} = \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1/3 \\ 0 \end{pmatrix}$, então $v_{33}=\dfrac{1}{9}$, $3v_{32}+4v_{33}=0 \Rightarrow v_{32} = -\dfrac{4}{27}$ e v_{31} qualquer $\Rightarrow \mathbf{v}_3 = \begin{pmatrix} 0 \\ -4/27 \\ 1/9 \end{pmatrix}$.

$\mathbf{V} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/3 & -4/27 \\ 0 & 0 & 1/9 \end{pmatrix} \Rightarrow \mathbf{V}^{-1}\,\mathbf{A}\,\mathbf{V} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} = \mathbf{J}.$

(b) $\mathbf{A} = \begin{pmatrix} 2 & 3 & 4 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \Rightarrow p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) = \det \begin{pmatrix} 2-\lambda & 3 & 4 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{pmatrix} = (2-\lambda)^3 \Rightarrow \lambda_1 = \lambda_2 = \lambda_3 = 2$

$\mathbf{A}-2\mathbf{I} = \begin{pmatrix} 0 & 3 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, posto$(\mathbf{A}-2\mathbf{I})=1$ então há dois vetores característicos associados a λ_1 que são determinados por:

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_1 = \begin{pmatrix} 3v_{12}+4v_{13} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. $(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_1 = \begin{pmatrix} 3v_{12}+4v_{13} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Então $v_{12}=4$, $v_{13}=-3$ e v_{11} qualquer $\Rightarrow \mathbf{v}_1 = \begin{pmatrix} 0 \\ 4 \\ -3 \end{pmatrix}$.

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_2 = \begin{pmatrix} 3v_{22}+4v_{23} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Então $v_{22}=v_{23}=0$ e $v_{21}\neq 0 \Rightarrow \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.

$(\mathbf{A}-2\mathbf{I})\,\mathbf{v}_3 = \begin{pmatrix} 3v_{32}+4v_{33} \\ 0 \\ 0 \end{pmatrix} = \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, então $v_{33}=0$, $v_{32}=\dfrac{1}{3}$ e v_{31} qualquer $\Rightarrow \mathbf{v}_3 = \begin{pmatrix} 0 \\ 1/3 \\ 0 \end{pmatrix}$.

$\mathbf{V} = \begin{pmatrix} 0 & 1 & 0 \\ 4 & 0 & 1/3 \\ -3 & 0 & 0 \end{pmatrix} \Rightarrow \mathbf{V}^{-1}\,\mathbf{A}\,\mathbf{V} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} = \mathbf{J}.$

(c) $\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \Rightarrow p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) = \det \begin{pmatrix} 2-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{pmatrix} = (2-\lambda)^3 \Rightarrow \lambda_1 = \lambda_2 = \lambda_3 = 2$

$\mathbf{A}-2\mathbf{I}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\end{pmatrix}$, posto$(\mathbf{A}-\lambda\mathbf{I})=0$ então há três vetores característicos associados a λ_1. Como o número máximo de vetores linearmente independentes de $\Re^3$ é igual à sua dimensão que é três, qualquer conjunto de três vetores linearmente independentes é composto por vetores característicos de $\mathbf{A}$, opta-se pela base canônica de $\Re^3$ resultando assim em $\mathbf{V}=\mathbf{I}$, pois, neste caso, a matriz $\mathbf{A}$ já está em sua forma canônica (neste caso uma matriz diagonal = $2\mathbf{I}$).

■

A.7 Formas Quadráticas

A expressão geral das formas quadráticas em $\Re^2$ é

$$f(x_1,x_2)=c+b_1x_1+b_2x_2+\frac{a_{11}}{2}x_1^2+a_{12}x_1x_2+\frac{a_{22}}{2}x_2^2,$$

cuja forma matricial é

$$f(\mathbf{x})=c+\mathbf{b}^T\,\mathbf{x}+\frac{1}{2}\mathbf{x}^T\,\mathbf{A}\,\mathbf{x},$$

sendo $\mathbf{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix}$, $\mathbf{b}=\begin{pmatrix}b_1\\b_2\end{pmatrix}$ e $\mathbf{A}=\begin{pmatrix}a_{11}&a_{12}\\a_{12}&a_{22}\end{pmatrix}$ (matriz simétrica).

Definindo o *operador diferencial gradiente* ∇ como o operador diferencial cujo componente i é $\nabla_i=\dfrac{\partial}{\partial x_i}$ e o *operador de Laplace*[6] ou *Laplaciano* (que é o *divergente* do vetor *gradiente*) por $\nabla^2=\nabla^T\,\nabla=\displaystyle\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2}$, que aplicados à função $f(\mathbf{x})$ acima, resulta em:

$$\nabla f(\mathbf{x})=\begin{pmatrix}\dfrac{\partial f(\mathbf{x})}{\partial x_1}\\[2ex]\dfrac{\partial f(\mathbf{x})}{\partial x_2}\end{pmatrix}=\mathbf{b}+\mathbf{A}\,\mathbf{x}\quad\text{e}\quad\nabla^2 f(\mathbf{x})=a_{11}+a_{22}=\text{tr}(\mathbf{A}).$$

Além destes operadores, define-se a *matriz Hessiana* de uma função escalar $f(\mathbf{x})$ como a matriz cujo elemento ij é $H_{ij}[f(\mathbf{x})]=\dfrac{\partial^2 f(\mathbf{x})}{\partial x_i\partial x_j}=\dfrac{\partial^2 f(\mathbf{x})}{\partial x_j\partial x_i}=H_{ji}[f(\mathbf{x})]$ (na realidade é a matriz Jacobiana do vetor gradiente de uma função escalar) , para a função $f(\mathbf{x})$ acima $\mathbf{H}[f(\mathbf{x})]=\mathbf{A}$.

Generalizando a expressão das formas quadráticas para $\Re^n$

$$f(\mathbf{x})=c+\sum_{i=1}^{n}b_ix_i+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j=c+\mathbf{b}^T\,\mathbf{x}+\frac{1}{2}\mathbf{x}^T\,\mathbf{A}\,\mathbf{x},$$

sendo $\mathbf{x}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}$, $\mathbf{b}=\begin{pmatrix}b_1\\b_2\\\vdots\\b_n\end{pmatrix}$, $\mathbf{A}=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{12}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&\vdots\!\vdots\!\vdots&\vdots\\a_{1n}&a_{2n}&\cdots&a_{nn}\end{pmatrix}$, $\nabla f(\mathbf{x})=\mathbf{b}+\mathbf{A}\,\mathbf{x}$,

$\nabla^2 f(\mathbf{x})=\text{tr}(\mathbf{A})$ e $\mathbf{H}[f(\mathbf{x})]=\mathbf{A}$.

Como $\mathbf{H}[f(\mathbf{x})]=\mathbf{A}$ e a matriz $\mathbf{H}$, por definição, é simétrica a matriz $\mathbf{A}$ também deve ser. Se uma matriz $\mathbf{Q}$ não for simétrica para torná-la simétrica basta fazer a transformação $\mathbf{A}\leftarrow\dfrac{1}{2}(\mathbf{Q}+\mathbf{Q}^T)$, pois $\mathbf{x}^T\,\mathbf{Q}\,\mathbf{x}=\mathbf{x}^T\,\mathbf{A}\,\mathbf{x}$.

[6] Pierre-Simon Laplace (1749-1827).

A forma quadrática acima pode ser simplificada, através de uma translação do eixo, eliminando o termo $\mathbf{b}^T\ \mathbf{x}$ assim, considerando $\mathbf{x} = \mathbf{y} + \mathbf{d}$, tem-se

$\mathbf{b}^T\ \mathbf{x} = \mathbf{b}^T\ \mathbf{y} + \mathbf{b}^T\ \mathbf{d}$

$\mathbf{x}^T\ \mathbf{A}\ \mathbf{x} = (\mathbf{y}^T + \mathbf{d}^T)\ (\mathbf{A}\ \mathbf{y} + \mathbf{A}\ \mathbf{d}) = \mathbf{y}^T\ \mathbf{A}\ \mathbf{y} + \mathbf{d}^T\ \mathbf{A}\ \mathbf{d} + \mathbf{y}^T\ \mathbf{A}\ \mathbf{d} + \mathbf{d}^T\ \mathbf{A}\ \mathbf{y}$, como $\mathbf{y}^T\ \mathbf{A}\ \mathbf{d} = \left(\mathbf{y}^T\ \mathbf{A}\ \mathbf{d}\right)^T = \mathbf{d}^T\ \mathbf{A}^T\ \mathbf{y} = \mathbf{d}^T\ \mathbf{A}\ \mathbf{y}$, pois $\mathbf{A}^T = \mathbf{A}$.

Assim $\mathbf{x}^T\ \mathbf{A}\ \mathbf{x} = (\mathbf{y}^T + \mathbf{d}^T)\ (\mathbf{A}\ \mathbf{y} + \mathbf{A}\ \mathbf{d}) = \mathbf{y}^T\ \mathbf{A}\ \mathbf{y} + \mathbf{d}^T\ \mathbf{A}\ \mathbf{d} + 2\mathbf{d}^T\ \mathbf{A}\ \mathbf{y}$ e

$$f(\mathbf{y}) = \left(c + \mathbf{b}^T\ \mathbf{d} + \frac{1}{2}\mathbf{d}^T\ \mathbf{A}\ \mathbf{d}\right) + (\mathbf{b} + \mathbf{A}\ \mathbf{d})^T\ \mathbf{y} + \frac{1}{2}\mathbf{y}^T\ \mathbf{A}\ \mathbf{y} = f(\mathbf{d}) +$$

$$+ (\mathbf{b} + \mathbf{A}\ \mathbf{d})^T\ \mathbf{y} + \frac{1}{2}\mathbf{y}^T\ \mathbf{A}\ \mathbf{y}.$$

Adotando $\mathbf{b} + \mathbf{A}\ \mathbf{d} = \mathbf{0} \Rightarrow \mathbf{d} = -\mathbf{A}^{-1}\ \mathbf{b}$ e $\hat{c} = f(\mathbf{d})$, resulta

$$f(\mathbf{y}) = \hat{c} + \frac{1}{2}\mathbf{y}^T\ \mathbf{A}\ \mathbf{y}.$$

Neste novo sistema de coordenadas tem-se $\nabla f(\mathbf{y}) = \begin{pmatrix} \dfrac{\partial f(\mathbf{y})}{\partial y_1} \\ \dfrac{\partial f(\mathbf{y})}{\partial y_2} \\ \vdots \\ \dfrac{\partial f(\mathbf{y})}{\partial y_n} \end{pmatrix} = \mathbf{A}\ \mathbf{y}$, assim, o valor da variável independente $\mathbf{y}$ que anula o vetor gradiente é o valor nulo $\mathbf{y} = \mathbf{0} \Rightarrow f(\mathbf{0}) = \hat{c}$. Deste modo, neste novo sistema de coordenadas, a origem $\mathbf{y} = \mathbf{0}$ é um *ponto crítico* que é uma condição necessária para o ponto ser um ponto de extremo (máximo ou mínimo) de $f(\mathbf{y})$. Se $\mathbf{y} = \mathbf{0}$ for um *ponto de mínimo* de $f(\mathbf{y})$, então para toda a vizinhança de $\mathbf{y} = \mathbf{0}$ em que $||\mathbf{y}|| \leq \delta$ deve-se ter $f(\mathbf{y}) > f(\mathbf{0}) = \hat{c} \Rightarrow \mathbf{y}^T\ \mathbf{A}\ \mathbf{y} > 0$ caracterizando a matriz $\mathbf{A}$ como *positiva definida*. Se $\mathbf{y} = \mathbf{0}$ for um *ponto de máximo* de $f(\mathbf{y})$ então para toda a vizinhança de $\mathbf{y} = \mathbf{0}$ em que $||\mathbf{y}|| \leq \delta$ deve-se ter $f(\mathbf{y}) < f(\mathbf{0}) = \hat{c} \Rightarrow \mathbf{y}^T\ \mathbf{A}\ \mathbf{y} < 0$ caracterizando a matriz $\mathbf{A}$ como *negativa definida*. Em qualquer outra situação o ponto não é nem de máximo ou mínimo, a matriz é dita ser *não definida* e o ponto crítico um *ponto de sela*.

A forma quadrática pode também ser rescrita em sua forma canônica, de forma análoga à apresentada no processo de transformação de matrizes à sua forma canônica, assim considerando $\mathbf{y} = \mathbf{V}\ \mathbf{z}$, em que $\mathbf{V}$ é a matriz cujos vetores coluna são os vetores característicos normalizados de $\mathbf{A}$ (considerados n vetores característicos linearmente independentes e ortogonais entre si, isto é, os valores característicos são todos reais pois a matriz $\mathbf{A}$ é simétrica), a matriz $\mathbf{V}$ é ortogonal, isto é, $\mathbf{V}^T\ \mathbf{V} = \mathbf{V}\ \mathbf{V}^T = \mathbf{I}$, então:

$$f(\mathbf{z}) = \hat{c} + \frac{1}{2}\mathbf{z}^T\ \left(\mathbf{V}^T\ \mathbf{A}\ \mathbf{V}\right)\ \mathbf{z} = \hat{c} + \frac{1}{2}\mathbf{z}^T\ \mathbf{D}\ \mathbf{z} = \hat{c} + \frac{1}{2}\sum_{i=1}^{n}\lambda_i z_i^2.$$

Como $\mathbf{y} = \mathbf{0}$ é um ponto crítico $\mathbf{z} = \mathbf{0}, \mathbf{z} = \mathbf{V}^{-1}\ \mathbf{y}$, é também um ponto crítico de $f(\mathbf{z})$, sendo um ponto de mínimo se todos os valores característicos de $\mathbf{A}$ forem positivos e um ponto de máximo se todos os valores característicos de $\mathbf{A}$ forem negativos, caso alguns valores característicos de $\mathbf{A}$ forem positivos e outros negativos o ponto $\mathbf{z} = \mathbf{0}$ $(\mathbf{y} = \mathbf{0})$, é um ponto de sela.

■ **Exemplo A.6** Análise dos pontos críticos em $\Re^2$.

(a) Ponto de Extremo (máximo ou mínimo) de $f(\mathbf{z})$ se λ_1 e λ_2 apresentarem o mesmo sinal, isto é, $\lambda_1\lambda_2 = \det(\mathbf{A}) > 0$. Sendo um *ponto de mínimo* se $\lambda_1 > 0$ e $\lambda_2 > 0$ e um *ponto de máximo* se $\lambda_1 < 0$ e $\lambda_2 < 0$. Em $\Re^3$ a superfície $f(\mathbf{z}) = \hat{c} + \lambda_1 z_1^2 + \lambda_2 z_2^2$ é uma *elipsoide* e no plano (z_1, z_2) as curvas $f(\mathbf{z}) = C$ são elipses com centro na origem. A Figura A.4 representa a superfície e as curvas de nível de $f(\mathbf{z}) = z_1^2 + 2z_2^2$, neste caso, $\mathbf{z} = \mathbf{0}$ é um ponto de mínimo no qual $f_{min}(\mathbf{z}) = 0$.

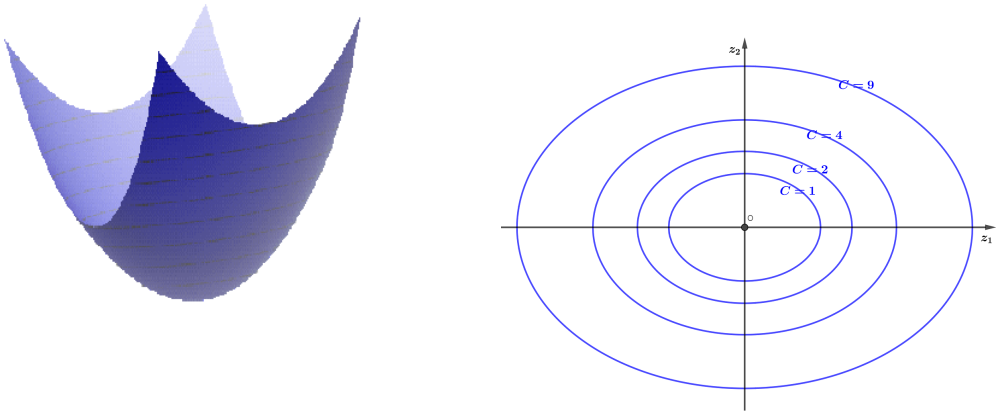


Figura A.4: Ponto de mínimo: elipsoide e elipses das curvas de nível.

(b) Ponto de Sela (nem máximo ou mínimo) de $f(\mathbf{z})$ se λ_1 e λ_2 apresentarem sinais distintos, isto é, $\lambda_1\lambda_2 = \det(\mathbf{A}) < 0$. Em $\Re^3$ a superfície $f(\mathbf{z}) = \hat{c} + \lambda_1 z_1^2 + \lambda_2 z_2^2$ é uma *hiperboloide* e no plano (z_1, z_2) as curvas $f(\mathbf{z}) = C$ são hipérboles. A Figura A.5 representa a superfície e as curvas de nível de $f(\mathbf{z}) = z_1^2 - 2z_2^2$, neste caso, $\mathbf{z} = \mathbf{0}$ é um ponto de sela no qual $f(\mathbf{z}) = 0$.

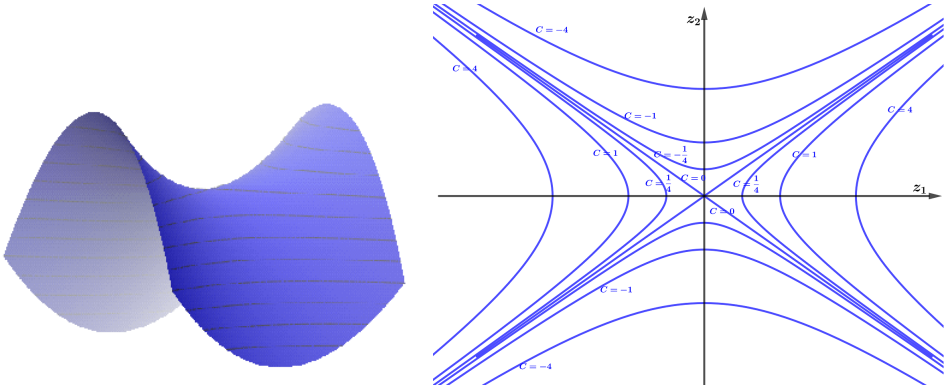


Figura A.5: Ponto de sela: hiperboloide e hipérboles das curvas de nível.

(c) Ponto Singular (insensível a variações em uma das direções) de $f(\mathbf{z})$ se $\lambda_1 = 0$ e $\lambda_2 \neq 0$ isto é, $\lambda_1\lambda_2 = \det(\mathbf{A}) = 0$, matriz $\mathbf{A}$ é singular. Como a matriz $\mathbf{A}$ é também simétrica, a única forma capaz de satisfazer a estas duas propriedades é $\mathbf{A} = \begin{pmatrix} a & a \\ a & a \end{pmatrix}$.
Além disto, para que
$f(x_1,x_2) = c + b_1x_1 + b_2x_2 + a\left(\frac{x_1^2}{2} + x_1x_2 + \frac{x_2^2}{2}\right) = c + b_1x_1 + b_2x_2 + \frac{a}{2}(x_1+x_2)^2$ tenha o gradiente nulo, é necessário que $\begin{pmatrix} a & a \\ a & a \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a\begin{pmatrix} x_1+x_2 \\ x_1+x_2 \end{pmatrix} = -\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \Rightarrow b_1 = b_2 = b$, resultando em $f(x_1,x_2) = c + b(x_1+x_2) + \frac{a}{2}(x_1+x_2)^2$. Que é a equação de uma parábola em (x_1+x_2) que apresenta um ponto de extremo em $x_1+x_2 = -\frac{b}{a}$, isto é, todos os pontos contidos nesta reta são *pontos críticos*, se $a > 0$ os pontos na reta são *pontos de mínimo* e se $a < 0$ os são *pontos de máximo*.

Os valores característicos de $\mathbf{A}$ são $\begin{cases}\lambda_1 = 0\\ \lambda_2 = 2a\end{cases}$ e os vetores característicos normalizados são $\begin{cases}\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\\ \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\end{cases}$.

A forma canônica de $\mathbf{A}$ é obtida pela transformação

$\mathbf{V}^T\,\mathbf{A}\,\mathbf{V} = \begin{pmatrix}0 & 0\\ 0 & 2a\end{pmatrix}$ em que $\mathbf{V} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}$.

A mudança da variável independente $\mathbf{x}$ para $\mathbf{z}$ é feita a partir de $\mathbf{x} = \mathbf{d} + \mathbf{V}\,\mathbf{z}$, em que $\mathbf{A}\,\mathbf{d} + \mathbf{b} = \mathbf{0} \Rightarrow a\begin{pmatrix}d_1+d_2\\ d_1+d_2\end{pmatrix} + b\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \Rightarrow d_1 + d_2 = -\frac{b}{a}$ e

$\hat{c} = f(d_1,d_2) = c - \frac{b^2}{2a}$, resultando na forma quadrática $f(\mathbf{z}) = \hat{c} + az_2^2$.

Em $\Re^3$ a superfície $f(\mathbf{z}) = \hat{c} + az_2^2$ é uma *paraboloide* e no plano (z_1, z_2) as curvas $f(\mathbf{z}) = C$ são retas paralelas ao eixo z_1, todos os pontos no eixo z_1 $(z_2 = 0)$ são pontos de mínimo se $a > 0$ e pontos de máximo se $a < 0$. A Figura A.6 representa a superfície e as curvas de nível para $f(\mathbf{z}) = z_2^2$, em que todos os pontos no eixo z_1 são pontos de mínimo (insensível a variações de z_1) e $f_{min}(\mathbf{z}) = 0$.

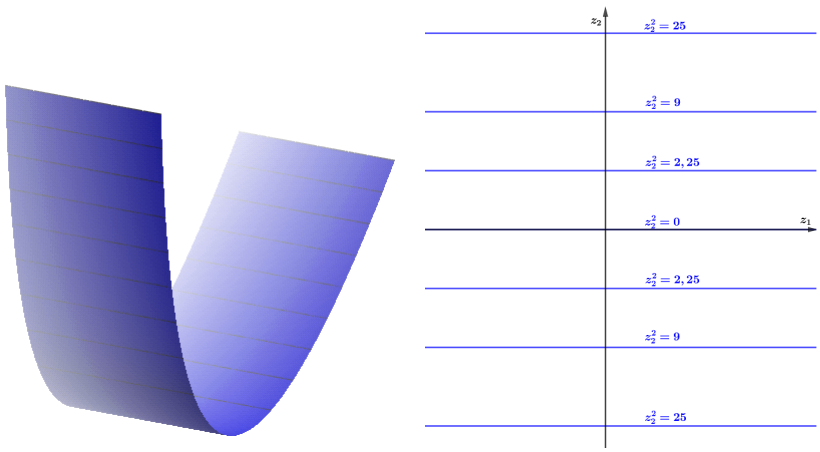


Figura A.6: Ponto singular: paraboloide e retas paralelas ao eixo z_1 como curvas de nível.

■

A.8 Funções de Matrizes

Funções escalares contínuas $f(x)$ com derivadas contínuas no intervalo $[a,b]$ e $x_0 \in [a,b]$, são ditas analíticas no intervalo e podem ser expressas por

$$f(x) = \sum_{k=0}^{\infty} c_k (x - x_0)^k \text{ em que } c_k = \frac{f^{(k)}(x_0)}{k!}.$$

Este conceito pode ser estendido a *funções de matrizes* de acordo com

$$f(\mathbf{A}) = \sum_{k=0}^{\infty} c_k \mathbf{A}^k.$$

Por analogia com a expansão $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, tem-se

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}.$$

Esta expansão apresenta as propriedades

(i) $\exp(\mathbf{0}) = \sum_{k=0}^{\infty} \frac{\mathbf{0}^k}{k!} = \mathbf{I}$.

(ii) Se t for um escalar então $\exp(\mathbf{A}t) = e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \left(\frac{t^k}{k!} \mathbf{A}^k \right)$, diferenciando esta expansão em relação a t, obtém-se

$\frac{d\left[\exp(\mathbf{A}t)\right]}{dt} = \frac{d\left(e^{\mathbf{A}t}\right)}{dt} = \sum_{k=1}^{\infty} \left(\frac{t^{k-1}}{(k-1)!} \mathbf{A}^k \right) = \mathbf{A} \sum_{k=0}^{\infty} \left(\frac{t^k}{k!} \mathbf{A}^k \right) = \mathbf{A}\, e^{\mathbf{A}t}$.

Adotando a notação $\Phi(t) = e^{\mathbf{A}t}$ tem-se $\begin{cases} \Phi(0) = \mathbf{I} \\ \frac{d\Phi(t)}{dt} = \mathbf{A}\,\Phi(t) \end{cases}$,

isto é, $\Phi(t)$ é solução da equação diferencial matricial linear

$\frac{d\Phi(t)}{dt} = \mathbf{A}\,\Phi(t)$ sujeita à condição inicial $\Phi(0) = \mathbf{I}$.

O Teorema de Cayley-Hamilton estabelece que

$\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + \cdots + c_2\mathbf{A}^2 + c_1\mathbf{A} + c_0\mathbf{I} = \mathbf{0}$, ou seja

$\mathbf{A}^n = -c_{n-1}\mathbf{A}^{n-1} + \cdots - c_2\mathbf{A}^2 - c_1\mathbf{A} - c_0\mathbf{I}$, multiplicando ambos os membros por $\mathbf{A}$

$\mathbf{A}^{n+1} = -c_{n-1}\mathbf{A}^n + \cdots - c_2\mathbf{A}^3 - c_1\mathbf{A}^2 - c_0\mathbf{A} = -c_{n-1}\left(-c_{n-1}\mathbf{A}^{n-1} + \cdots - c_2\mathbf{A}^2 - c_1\mathbf{A} - c_0\mathbf{I}\right) + \cdots - c_2\mathbf{A}^3 - c_1\mathbf{A}^2 - c_0\mathbf{A} = \alpha_{n-1}\mathbf{A}^{n-1} + \cdots + \alpha_2\mathbf{A}^2 + \alpha_1\mathbf{A} + \alpha_0\mathbf{I}$ repetindo a operação

$\mathbf{A}^{n+2} = \beta_{n-1}\mathbf{A}^{n-1} + \cdots + \beta_2\mathbf{A}^2 + \beta_1\mathbf{A} + \beta_0\mathbf{I}$.

E assim sucessivamente, permitindo concluir que

$$\mathbf{A}^m = \gamma_{n-1}\mathbf{A}^{n-1} + \gamma_{n-2}\mathbf{A}^{n-2} + \cdots + \gamma_2\mathbf{A}^2 + \gamma_1\mathbf{A} + \gamma_0\mathbf{I},$$

expressão válida para $m = 0, 1, 2, 3, \cdots$ e se $\mathbf{A}$ for não singular para $m = 0, \pm1, \pm2, \pm3, \cdots$.

Aplicando esta expressão nas expansões de funções de matrizes

$$f(\mathbf{A}) = \sum_{k=0}^{\infty} c_k\mathbf{A}^k = \sum_{k=0}^{n-1} \gamma_k\mathbf{A}^k.$$

Como o Teorema de Cayley-Hamilton estabelece que as funções polinomiais dos valores característicos da matriz *valem* também para a matriz, pode-se afirmar a recíproca: *funções polinomiais da matriz são também atendidas por seus valores característicos*.

Assim, os coeficientes de $f(\mathbf{A}) = \sum_{k=0}^{n-1} \gamma_k\mathbf{A}^k$, pode ser determinados pela resolução do sistema linear de equações:

$\sum_{k=0}^{n-1} \gamma_k\lambda_j^k = f(\lambda_j)$, para $j = 1, 2, \cdots, n$, se os valores característicos de $\mathbf{A}$ forem distintos.

Por analogia com a interpolação polinomial de Lagrange, Sylvester[7] propôs a seguinte fórmula (*Fórmula de Sylvester*) para calcular funções de matrizes:

$$f(\mathbf{A}) = \sum_{i=1}^{n} f(\lambda_i) \left[\prod_{k=1\neq i}^{n} \left(\frac{\mathbf{A} - \lambda_k\mathbf{I}}{\lambda_i - \lambda_k} \right) \right], \text{ se os valores característicos de } \mathbf{A} \text{ forem distintos.}$$

[7] James Joseph Sylvester (1814-1897).

Se $\mathbf{A}$ for uma matriz (2,2), tem-se $\begin{cases}\gamma_0+\gamma_1\lambda_1=f(\lambda_1)\\ \gamma_0+\gamma_1\lambda_2=f(\lambda_2)\end{cases}\Rightarrow\begin{cases}\gamma_0=\dfrac{\lambda_2 f(\lambda_1)-\lambda_1 f(\lambda_2)}{\lambda_2-\lambda_1}\\ \gamma_1=\dfrac{f(\lambda_2)-f(\lambda_1)}{\lambda_2-\lambda_1}\end{cases}$

então $f(\mathbf{A})=\left(\dfrac{\lambda_2 f(\lambda_1)-\lambda_1 f(\lambda_2)}{\lambda_2-\lambda_1}\right)\mathbf{I}+\left(\dfrac{f(\lambda_2)-f(\lambda_1)}{\lambda_2-\lambda_1}\right)\mathbf{A}$,

ou pela fórmula de Sylvester $f(\mathbf{A})=f(\lambda_1)\dfrac{\mathbf{A}-\lambda_2\mathbf{I}}{\lambda_1-\lambda_2}+f(\lambda_2)\dfrac{\mathbf{A}-\lambda_1\mathbf{I}}{\lambda_2-\lambda_1}$.

Caso $\lambda_1=\lambda_2\Rightarrow\begin{cases}\gamma_0=\lim\limits_{\lambda_2\to\lambda_1}\left(\dfrac{\lambda_2 f(\lambda_1)-\lambda_1 f(\lambda_2)}{\lambda_2-\lambda_1}\right)=f(\lambda_1)\\ \gamma_1=\lim\limits_{\lambda_2\to\lambda_1}\left(\dfrac{f(\lambda_2)-f(\lambda_1)}{\lambda_2-\lambda_1}\right)=f'(\lambda_1)\end{cases}$, então

$f(\mathbf{A})=f(\lambda_1)\mathbf{I}+f'(\lambda_1)\mathbf{A}$. Resultado análogo obtém-se aplicando o limite $\lambda_2\to\lambda_1$ na fórmula de Sylvester.

Estes procedimentos podem ser estendidos para funções $f(\mathbf{A}t)=\sum\limits_{k=0}^{n-1}\gamma_k(t)\mathbf{A}^k$, em que $\gamma_k(t)$ são funções da variável escalar t determinadas por:

$\sum\limits_{k=0}^{n-1}\gamma_k(t)\lambda_j^k=f(\lambda_j t)$, para $j=1,2,\cdots,n$, ou pela fórmula de Sylvester:

$f(\mathbf{A}t)=\sum\limits_{i=1}^{n}f(\lambda_i t)\left[\prod\limits_{k=1\neq i}^{n}\left(\dfrac{\mathbf{A}-\lambda_k\mathbf{I}}{\lambda_i-\lambda_k}\right)\right]$.

Ambas expressões válidas apenas se os valores característicos de $\mathbf{A}$ forem distintos.

Caso a matriz $\mathbf{A}$ apresentar valores característicos múltiplos, os procedimentos apresentados devem ser modificados. Por exemplo, considera-se $\lambda_1=\lambda_2=\cdots=\lambda_m\neq\lambda_{m+1}\neq\lambda_{m+2}\neq\cdots\neq\lambda_n$, os coeficientes de $f(\mathbf{A})=\sum\limits_{k=0}^{n-1}\gamma_k\mathbf{A}^k$, são então determinados pela resolução do sistema linear de equações:

$$\begin{cases}\sum\limits_{k=0}^{n-1}\gamma_k\lambda_1^k=f(\lambda_1)\\ \sum\limits_{k=1}^{n-1}k\gamma_k\lambda_1^{k-1}=f'(\lambda_1)\\ \sum\limits_{k=2}^{n-1}k(k-1)\gamma_k\lambda_1^{k-2}=f\text{''}(\lambda_1)\\ \vdots\vdots\vdots\\ \sum\limits_{k=m}^{n-1}k(k-1)\cdots(k-m-1)\gamma_k\lambda_1^{k-m}=f^{(m)}(\lambda_1)\\ \sum\limits_{k=0}^{n-1}\gamma_k(t)\lambda_j^k=f(\lambda_j)\text{ para }j=(m+1),(m+2),\cdots,n\end{cases},$$

ou pela *fórmula de Sylvester modificada*:

$f(\mathbf{A})=\sum\limits_{i=0}^{m}f^{(i)}(\lambda_1)\dfrac{(\mathbf{A}-\lambda_1\mathbf{I})^i}{i!}+\dfrac{(\mathbf{A}-\lambda_1\mathbf{I})^{m+1}}{(m+1)!}\sum\limits_{i=m+1}^{n}g(\lambda_i)\left[\prod\limits_{k=1\neq i}^{n}\left(\dfrac{\mathbf{A}-\lambda_k\mathbf{I}}{\lambda_i-\lambda_k}\right)\right]$ em que $g(\lambda)=\dfrac{(m+1)!}{(\lambda-\lambda_1)^{m+1}}\left[f(\lambda)-\sum\limits_{i=0}^{m}f^{(i)}(\lambda_1)\dfrac{(\lambda-\lambda_1)^i}{i!}\right]$.

Neste caso, para o cálculo de $f(\mathbf{A}t)=\sum\limits_{k=0}^{n-1}\gamma_k(t)\mathbf{A}^k$, assim se procede:

$$\begin{cases} \sum_{k=0}^{n-1} \gamma_k(t)\lambda_1^k = f(\lambda_1 t) \\ \sum_{k=1}^{n-1} k\gamma_k(t)\lambda_1^{k-1} = \left.\dfrac{\partial f(\lambda t)}{\partial \lambda}\right|_{\lambda=\lambda_1} \\ \sum_{k=2}^{n-1} k(k-1)\gamma_k(k)\lambda_1^{k-2} = \left.\dfrac{\partial^2 f(\lambda t)}{\partial \lambda^2}\right|_{\lambda=\lambda_1} \\ \vdots\vdots\vdots \\ \sum_{k=m}^{n-1} k(k-1)\cdots(k-m-1)\gamma_k(t)\lambda_1^{k-m} = \left.\dfrac{\partial^m f(\lambda t)}{\partial \lambda^m}\right|_{\lambda=\lambda_1} \\ \\ \sum_{k=0}^{n-1} \gamma_k(t)\lambda_j^k = f(\lambda_j) \text{ para } j=(m+1),(m+2),\cdots,n \end{cases},$$

ou pela fórmula de Sylvester modificada:

$$f(\mathbf{A}t) = \sum_{i=0}^{m} \left.\frac{\partial^i f(\lambda t)}{\partial \lambda^i}\right|_{\lambda=\lambda_1} \frac{(\mathbf{A}-\lambda_1\mathbf{I})^i}{i!} + \frac{(\mathbf{A}-\lambda_1\mathbf{I})^{m+1}}{(m+1)!} \sum_{i=m+1}^{n} g(\lambda_i t)\left[\prod_{k=1\neq i}^{n}\left(\frac{\mathbf{A}-\lambda_k\mathbf{I}}{\lambda_i-\lambda_k}\right)\right] \text{ em que}$$

$$g(\lambda t) = \frac{(m+1)!}{(\lambda-\lambda_1)^{m+1}}\left[f(\lambda) - \sum_{i=0}^{m} \left.\frac{\partial^i f(\lambda t)}{\partial \lambda^i}\right|_{\lambda=\lambda_1} \frac{(\lambda-\lambda_1)^i}{i!}\right].$$

■ **Exemplo A.7** Para ilustrar o emprego das expressões anteriores para o cálculo de funções de matrizes, vários exemplos são a seguir apresentados.

1. Cálculo de $\mathbf{A}^{-3}$ e $\ln(\mathbf{A})$ da matriz $\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$.

   $p(\lambda) = \lambda^2 - 7\lambda + 10 = (\lambda-2)(\lambda-5) \Rightarrow \lambda_1 = 2$ e $\lambda_2 = 5$, $\mathbf{A}-\lambda_1\mathbf{I} = \begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$ e

   $\mathbf{A}-\lambda_2\mathbf{I} = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}$.

   $f(\mathbf{A}) = \left(\dfrac{5f(2)-2f(5)}{3}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \left(\dfrac{f(5)-f(2)}{3}\right)\begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$,

   ou pela fórmula de Sylvester $f(\mathbf{A}) = -\dfrac{f(2)}{3}\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} + \dfrac{f(5)}{3}\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$.

   Cálculo de $\mathbf{A}^{-3} \Rightarrow f(2) = \dfrac{1}{2^3}$ e $f(5) = \dfrac{1}{5^3}$, logo

   $f(\mathbf{A}) = \dfrac{5/2^3 - 2/5^3}{3}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \dfrac{1/5^3 - 1/2^3}{3}\begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix} = \dfrac{1}{1000}\begin{pmatrix} 47 & -78 \\ -39 & 86 \end{pmatrix}$.

   Pela fórmula de Sylvester $f(\mathbf{A}) = \dfrac{1/5^3}{3}\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix} - \dfrac{1/2^3}{3}\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} = \dfrac{1}{1000}\begin{pmatrix} 47 & -78 \\ -39 & 86 \end{pmatrix}$.

   Cálculo de $\ln(\mathbf{A}) \Rightarrow f(2) = \ln(2)$ e $f(5) = \ln(5)$, logo

   $f(\mathbf{A}) = \left(\dfrac{5\ln(2)-2\ln(5)}{3}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \left(\dfrac{\ln(5)-\ln(2)}{3}\right)\begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix} = \dfrac{1}{3}\begin{pmatrix} \ln(50) & \ln(25/4) \\ \ln(5/2) & \ln(20) \end{pmatrix}$.

   Pela Fórmula de Sylvester $f(\mathbf{A}) = \dfrac{\ln(5)}{3}\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix} - \dfrac{\ln(2)}{3}\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} = \dfrac{1}{3}\begin{pmatrix} \ln(50) & \ln(25/4) \\ \ln(5/2) & \ln(20) \end{pmatrix}$.

2. Cálculo de $\sqrt{\mathbf{A}}$ da matriz $\mathbf{A} = \begin{pmatrix} 5 & -3 \\ 2 & -2 \end{pmatrix}$.

   $p(\lambda) = \lambda^2 - 3\lambda - 4 = (\lambda+1)(\lambda-4) \Rightarrow \lambda_1 = -1$ e $\lambda_2 = 4$, $\mathbf{A}-\lambda_1\mathbf{I} = \begin{pmatrix} 6 & -3 \\ 2 & -1 \end{pmatrix}$ e

   $\mathbf{A}-\lambda_2\mathbf{I} = \begin{pmatrix} 1 & -3 \\ 2 & -6 \end{pmatrix}$.

Cálculo de $f(\lambda_1)=f(-1)=\mathbf{i}$ e $f(\lambda_2)=f(4)=2$, logo
$f(\mathbf{A})=\left(\frac{4\mathbf{i}+2}{5}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix}+\left(\frac{2-\mathbf{i}}{5}\right)\begin{pmatrix}5&-3\\2&-2\end{pmatrix}=\frac{1}{5}\begin{pmatrix}12-\mathbf{i}&-6+3\mathbf{i}\\4-2\mathbf{i}&-2+6\mathbf{i}\end{pmatrix}$.
Pela fórmula de Sylvester $f(\mathbf{A})=-\frac{\mathbf{i}}{5}\begin{pmatrix}1&-3\\2&-6\end{pmatrix}+\frac{2}{5}\begin{pmatrix}6&-3\\2&-1\end{pmatrix}=\frac{1}{5}\begin{pmatrix}12-\mathbf{i}&-6+3\mathbf{i}\\4-2\mathbf{i}&-2+6\mathbf{i}\end{pmatrix}$.
Verificando-se que $\left[\frac{1}{5}\begin{pmatrix}12-\mathbf{i}&-6+3\mathbf{i}\\4-2\mathbf{i}&-2+6\mathbf{i}\end{pmatrix}\right]^2=\begin{pmatrix}5&-3\\2&-2\end{pmatrix}=\mathbf{A}$.

3. Cálculo de $\exp(\mathbf{A})$ da matriz $\mathbf{A}=\begin{pmatrix}-1,75&-2&-0,250\\-0,125&-2&-0,375\\0,250&2&-0,250\end{pmatrix}$.
$p(\lambda)=\lambda^3+4\lambda^2+5\lambda+2=(\lambda+1)^2(\lambda+2)\Rightarrow\lambda_1=\lambda_2=-1$ e $\lambda_3=-2$, $\mathbf{A}-\lambda_1\mathbf{I}=\begin{pmatrix}-0,75&-2&-0,250\\-0,125&-1&-0,375\\0,250&2&0,750\end{pmatrix}$ e
$(\mathbf{A}-\lambda_1\mathbf{I})^2=\begin{pmatrix}0,750&3&0,750\\0,125&0,5&0,125\\-0,25&-1&-0,250\end{pmatrix}$.
Cálculo de $f(\lambda_1)=f(-1)=e^{-1}$, $f'(\lambda_1)=f'(-1)=e^{-1}$ e $f(\lambda_3)=f(2)=e^{-2}$,
$\begin{cases}\gamma_0-\gamma_1+\gamma_2=e^{-1}\\\gamma_1-2\gamma_2=e^{-1}\\\gamma_0-2\gamma_1+4\gamma_2=e^{-2}\end{cases}\Rightarrow\begin{cases}\gamma_0=0,871094\\\gamma_1=0,638550\\\gamma_2=0,135335\end{cases}$, logo
$f(\mathbf{A})=0,871094\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}+0,638550\begin{pmatrix}-1,75&-2&-0,250\\-0,125&-2&-0,375\\0,250&2&-0,250\end{pmatrix}+$
$+0,135335\left[\begin{pmatrix}-1,75&-2&-0,250\\-0,125&-2&-0,375\\0,250&2&-0,250\end{pmatrix}\right]^2=\begin{pmatrix}0,193471&-0,329753&0,009532\\-0,029068&0,067668&-0,121038\\0,058136&0,600424&0,609955\end{pmatrix}$.
Pela fórmula de Sylvester com $g(\lambda_3)=\frac{2}{(-1)^2}\left[e^{-2}-e^{-1}(1-1)\right]=2e^{-2}$
$f(\mathbf{A})=e^{-1}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}+e^{-1}\begin{pmatrix}-0,75&-2&-0,250\\-0,125&-1&-0,375\\0,250&2&0,750\end{pmatrix}+e^{-2}\begin{pmatrix}0,750&3&0,750\\0,125&0,5&0,125\\-0,25&-1&-0,250\end{pmatrix}=$
$=\begin{pmatrix}0,193471&-0,329753&0,009532\\-0,029068&0,067668&-0,121038\\0,058136&0,600424&0,609955\end{pmatrix}$.

4. Cálculo de $\exp(\mathbf{A}t)$ da matriz $\mathbf{A}=\begin{pmatrix}4&2\\1&3\end{pmatrix}$.
$p(\lambda)=\lambda^2-7\lambda+10=(\lambda-2)(\lambda-5)\Rightarrow\lambda_1=2$ e $\lambda_2=5$, $\mathbf{A}-\lambda_1\mathbf{I}=\begin{pmatrix}2&2\\1&1\end{pmatrix}$ e
$\mathbf{A}-\lambda_2\mathbf{I}=\begin{pmatrix}-1&2\\1&-2\end{pmatrix}$.
Cálculo de $f(\lambda_1 t)=e^{2t}$ e $f(\lambda_2 t)=f(4)=e^{5t}$, logo
$\exp(\mathbf{A}t)=\left(\frac{5e^{2t}-2e^{5t}}{3}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix}+\left(\frac{e^{5t}-e^{2t}}{3}\right)\begin{pmatrix}4&2\\1&3\end{pmatrix}=\frac{1}{3}\begin{pmatrix}2e^{5t}+e^{2t}&2(e^{5t}-e^{2t})\\e^{5t}-e^{2t}&e^{5t}+2e^{2t}\end{pmatrix}$.
Pela fórmula de Sylvester:
$\exp(\mathbf{A}t)=-\frac{e^{2t}}{3}\begin{pmatrix}-1&2\\1&-2\end{pmatrix}+\frac{e^{5t}}{3}\begin{pmatrix}2&2\\1&1\end{pmatrix}=\frac{1}{3}\begin{pmatrix}2e^{5t}+e^{2t}&2(e^{5t}-e^{2t})\\e^{5t}-e^{2t}&e^{5t}+2e^{2t}\end{pmatrix}$.

Verificando-se que
$\frac{d[\exp(\mathbf{A}t)]}{dt} = \frac{d}{dt}\left[\frac{1}{3}\begin{pmatrix} 2e^{5t}+e^{2t} & 2(e^{5t}-e^{2t}) \\ e^{5t}-e^{2t} & e^{5t}+2e^{2t}\end{pmatrix}\right] = \frac{1}{3}\begin{pmatrix} 10e^{5t}+2e^{2t} & 2(5e^{5t}-2e^{2t}) \\ 5e^{5t}-2e^{2t} & 5e^{5t}+4e^{2t}\end{pmatrix} =$
$= \begin{pmatrix} 4 & 2 \\ 1 & 3\end{pmatrix}\left[\frac{1}{3}\begin{pmatrix} 2e^{5t}+e^{2t} & 2(e^{5t}-e^{2t}) \\ e^{5t}-e^{2t} & e^{5t}+2e^{2t}\end{pmatrix}\right] = \mathbf{A}\exp(\mathbf{A}t)$ e $\exp(\mathbf{A}t)|_{t=0} = \mathbf{I}$.

5. Cálculo de $\exp(\mathbf{A}t)$ da matriz $\mathbf{A} = \begin{pmatrix} -3 & -2 \\ 2 & 1\end{pmatrix}$.

$p(\lambda) = \lambda^2 + 2\lambda + 1 = (\lambda+1)^2 \Rightarrow \lambda_1 = \lambda_2 = -1$.
Cálculo de $f(\lambda_1 t) = e^{-t}$ e $\left.\frac{\partial e^{\lambda t}}{\partial \lambda}\right|_{\lambda=\lambda_1} = te^{-t}$, logo
$\exp(\mathbf{A}t) = e^{-t}\begin{pmatrix} 1 & 0 \\ 0 & 1\end{pmatrix} + te^{-t}\begin{pmatrix} -3 & -2 \\ 2 & 1\end{pmatrix} = e^{-t}\begin{pmatrix} 1-2t & -2t \\ 2t & 1+2t\end{pmatrix}$.
Procedimento análogo pela fórmula de Sylvester.
Verificando-se que
$\frac{d[\exp(\mathbf{A}t)]}{dt} = \frac{d}{dt}\left[e^{-t}\begin{pmatrix} 1-2t & -2t \\ 2t & 1+2t\end{pmatrix}\right] = e^{-t}\begin{pmatrix} 2t-3 & 2(t-1) \\ 2(1-t) & 3-2t\end{pmatrix} =$
$= \begin{pmatrix} -3 & -2 \\ 2 & 1\end{pmatrix}\left[e^{-t}\begin{pmatrix} 1-2t & -2t \\ 2t & 1+2t\end{pmatrix}\right] = \mathbf{A}\exp(\mathbf{A}t)$ e $\exp(\mathbf{A}t)|_{t=0} = \mathbf{I}$.

■

A.9 Sistemas de Equações Diferenciais Lineares

Equações diferenciais lineares de primeira ordem da forma:
$\frac{dx(t)}{dt} = a(t)x(t) + b(t)u(t)$ para $t > t_0$, com $x(t_0) = x_0$ (condição inicial),
sendo $x(t)$ a *variável de saída* ou *variável de estado*, $u(t)$ a *variável de entrada* ou *perturbação* e $a(t)$ e $b(t)$ os *parâmetros* do problema (funções conhecidas da variável independente t), apresentam a solução $x(t) = x_h(t) + x_p(t)$ para $t > t_0$, em que:

$$\begin{cases} \text{Solução homogênea: } \frac{dx_h(t)}{dt} = a(t)x_h(t) \text{ com } x(t_0) = x_0 \\ \text{Solução particular: } \frac{dx_p(t)}{dt} = a(t)x_p(t) + b(t)u(t) \text{ com } x_p(t_0) = 0 \end{cases}.$$

A solução homogênea pode ser obtida por integração direta da equação, resultando em

$$x_h(t) = \phi(t;t_0)x_0, \text{ sendo } \phi(t;t_0) = \exp\left[\int_{\xi=t_0}^{t} a(\xi)d\xi\right]$$

Na realidade, $\phi(t;t_0)$ é solução da equação diferencial homogênea com a condição inicial unitária, isto é, $\frac{d\phi(t;t_0)}{dt} = a(t)\phi(t;t_0)$ com $\phi(t_0;t_0) = 1$.

Para determinar a solução particular, aplica-se o método de *variação de parâmetros*, que consiste em buscar a solução da forma $x_p(t) = \phi(t;t_0)z(t)$ com $z(t_0) = 0$, e, em vista de $\frac{dx_p(t)}{dt} = \frac{d\phi(t;t_0)}{dt}z(t) + \phi(t;t_0)\frac{dz(t)}{dt} = \phi(t;t_0)\left[\frac{dz(t)}{dt} + a(t)z(t)\right]$, ou seja,
$a(t)x_p(t) + \phi(t;t_0)\frac{dz(t)}{dt} = a(t)x_p(t) + b(t)u(t) \Rightarrow \frac{dz(t)}{dt} = \frac{b(t)u(t)}{\phi(t;t_0)}$ com $z(t_0) = 0$, obtendo-se por integração $z(t) = \int_{\xi=t_0}^{t}\frac{b(\xi)u(\xi)}{\phi(\xi;t_0)}d\xi \Rightarrow x_p(t) = \phi(t;t_0)\left[\int_{\xi=t_0}^{t}\frac{b(\xi)u(\xi)}{\phi(\xi;t_0)}d\xi\right]$. A solução geral

do problema é:

$$x(t) = \phi(t;t_0)x_0 + \phi(t;t_0)\left[\int_{\xi=t_0}^{t} \frac{b(\xi)u(\xi)}{\phi(\xi;t_0)}d\xi\right], \text{ sendo } \phi(t;t_0) = \exp\left[\int_{\xi=t_0}^{t} a(\xi)d\xi\right].$$

Estrutura equivalente aplica-se a um sistema de n equações diferenciais lineares expresso na forma:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{u}(t) \text{ para } t > t_0, \text{ com } \mathbf{x}(t_0) = \mathbf{x}_0.$$

em que $\mathbf{x}(t) \in \Re^n$, $\mathbf{u}(t) \in \Re^m$, $\mathbf{A}(t) \in \Re^{n\times n}$ e $\mathbf{A}(t) \in \Re^{n\times m}$.

A solução do sistema pode ser expressa por (de maneira semelhante ao caso escalar) $\mathbf{x}(t) = \mathbf{x}_h(t) + \mathbf{x}_p(t)$, expressando $\mathbf{x}_h(t) = \Phi(t;t_0)\,\mathbf{x}_0$ e $\mathbf{x}_p(t) = \Phi(t;t_0)\,\mathbf{z}(t)$ sendo $\Phi(t;t_0) \in \Re^{n\times n}$ a *matriz de transição* do sistema, solução da equação diferencial matricial:

$$\frac{d\Phi(t;t_0)}{dt} = \mathbf{A}(t)\,\Phi(t;t_0) \text{ com } \Phi(t_0;t_0) = \mathbf{I}.$$

A função $\mathbf{z}(t)$ em $\mathbf{x}_p(t) = \Phi(t;t_0)\,\mathbf{z}(t)$ é determinada substituindo esta expressão em $\frac{d\mathbf{x}_p(t)}{dt} = \mathbf{A}(t)\,\mathbf{x}_p(t) + \mathbf{B}(t)\,\mathbf{u}(t)$ com $\mathbf{x}_p(t_0) = \mathbf{0}$, dando origem a $\Phi(t;t_0)\,\frac{d\mathbf{z}(t)}{dt} = \mathbf{B}(t)\,\mathbf{u}(t)$ com $\mathbf{z}(t_0) = \mathbf{0}$, cuja solução é:

$$\mathbf{z}(t) = \int_{\xi=t_0}^{t} [\Phi(\xi;t_0)]^{-1}\,\mathbf{B}(\xi)\,\mathbf{u}(\xi)d\xi \Rightarrow \mathbf{x}_p(t) = \Phi(t;t_0)\,\left[\int_{\xi=t_0}^{t} [\Phi(\xi;t_0)]^{-1}\,\mathbf{B}(\xi)\,\mathbf{u}(\xi)d\xi\right].$$

Desse modo, a solução geral do problema é :

$$\mathbf{x}(t) = \Phi(t;t_0)\,\mathbf{x}_0 + \Phi(t;t_0)\,\left[\int_{\xi=t_0}^{t} [\Phi(t;t_0)]^{-1}\,\mathbf{B}(\xi)\,\mathbf{u}(\xi)d\xi\right],$$

em que a *matriz de transição* $\Phi(t;t_0)$ é solução da equação diferencial matricial:

$$\frac{d\Phi(t;t_0)}{dt} = \mathbf{A}(t)\,\Phi(t;t_0) \text{ com } \Phi(t_0;t_0) = \mathbf{I}.$$

Se a matriz $\mathbf{A}(t)$ for constante $\mathbf{A}(t) = \mathbf{A}$, a *matriz de transição* é $\Phi(t;t_0) = \exp[\mathbf{A}(t-t_0)]$ e a solução geral do sistema é:

$$\mathbf{x}(t) = \exp[\mathbf{A}(t-t_0)]\,\mathbf{x}_0 + \int_{\xi=t_0}^{t} \exp[\mathbf{A}(t-\xi)]\,\mathbf{B}(\xi)\,\mathbf{u}(\xi)d\xi.$$

Bibliografia

Artigos

Broyden, C.G. (1965). "A Class of Methods for Solving Nonlinear Simultaneous Equations". Em: *Mathematics of Computation* 19, páginas 577–593 (ver página 146).

Bulirsch, Roland e Josef Stoer (1966). "Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods". Em: *Numerische Mathematik* 8, páginas 1–13 (ver página 54).

Butcher, John C. (1996). "A History of Runge-Kutta Methods". Em: *Applied Numerical Mathematics* 20, páginas 247–260 (ver página 210).

Clenshaw, C. W. (1955). "A Note on the Summation of Chebyshev Series". Em: *Mathematics of Computation* 9.51, páginas 118–120. DOI: 10.1090/S0025-5718-1955-0071856-0 (ver página 70).

Coggins, G.F. (1964). "Univariate Search Method". Em: *Imperial Chemical Industry, Central Instrument Laboratory Research Note* 64, página 11 (ver página 243).

Gill, S. (1951). "A Process for the Step-by-Step Integration of Differential Equations in an Automatic Digital Computing Machine". Em: *Mathematical Proceedings of the Cambridge Philosophical Society* 47, páginas 96–108 (ver página 211).

Hooke, R. e T.A. Jeeves (1961). "Direct Search Solution of Numerical and Statistical Problems". Em: *J. ACM* 8, páginas 212–220 (ver página 243).

Johnson, G.E. e M.A. Townsend (1978). "Nonoptimal Termination Properties of Quadratic Interpolation Univariate Searches". Em: *Journal of the Franklin Institute* 306.3, páginas 257–266 (ver página 243).

Kennedy, James e Russell Eberhart (1995). "Particle Swarm Optimization". Em: *In Proceedings of the IV IEEE International Conference on Neural Networks* 1, 1942–1948 (ver página 246).

Nelder, J.A. e R. Mead (1965). "A Simplex Method for Function Minimization". Em: *The Computer Journal* 7, páginas 308–320 (ver página 245).

Sherman, Jack e Winifred J. Morrison (1950). "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix". Em: *Annals of Mathematical Statistics* 21.1, páginas 124–127 (ver página 137).

Weisstein, Eric W. (2006). "Neville's Algorithm". Em: *MathWorld: NevillesAlgorithm.html*, (último acesso em 02/2020) (ver página 54).

Livros

Aris, Rutherford (1999). *Mathematical Modeling - A Chemical Engineer's Perspective*. 1ª edição. New York: Academic Press (ver página 75).

Brenan, K.E., S.L. Campbell e L.R. Petzold (1996). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. 1ª edição. Philadelphia: SIAM (ver página 223).

Butcher, John C. (1987). *The Numerical Analysis of Ordinary Differential Equations – Runge-Kutta and General Linear Methods*. 1ª edição. New York: Wiley (ver página 211).

Carnahan, B., H.A. Luther e J.O. Wilkes (1969). *Applied Numerical Methods*. 1ª edição. New York: John Wiley & Sons, Inc. (ver página 152).

Edgar, Thomas F., David M. Himmelblau e Leon S. Lasdon (2001). *Optimization of Chemical Processes*. 2ª edição. New York: McGraw-Hill (ver página 240).

Hougen, Olaf A. e Kenneth M. Watson (1955). *Chemical Process Principles*. 1ª edição. New York: John Wiley & Sons, Inc. (ver página 71).

Marcus, Marvin (1960). *Basic Theorems in Matrix Theory*. 1ª edição. Volume 57. New York: Applied Mathematics Series, National Bureau of Standards (ver página 124).

Perlingeiro, Carlos Augusto G. (2005). *Engenharia de Processos: Análise, Simulação, Otimização e Síntese de Processos Químicos*. 1ª edição. Rio de Janeiro: Blucher (ver página 229).

Spiegel, Murray R. e John Liu (1999). *Mathematical Handbook of Formulas and Tables*. 1ª edição. New York: McGraww-Hill, Schaum's Outline Series (ver páginas 186, 187).

Swann, W.H. (1972). *Direct Search Methods*. 1ª edição. New York: W. Murray (Ed.), Em Numerical Methods for Unconstrained Optimization, Academic Press (ver página 243).

Livros Complementares

Amundson, Neal R. (1966). *Mathematical Methods in Chemical Engineering: Matrices and Their Application*. 1ª edição. New Jersey: Prenticel Hall, Inc. (ver página 5).

Ascher, Uri M. e Linda R. Petzold (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. 1ª edição. Philadelphia: SIAM.

Biegler, Lorenz T. (2010). *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. 1ª edição. Philadelphia: SIAM e MOS.

Burden, Richard e J. Douglas Faires (2003). *Análise Numérica*. 1ª edição. Toronto: Thomson.

Conte, Samuel D. e Carl de Boor (1981). *Elementary Numerical Analysis: An Algortithm Approach*. 1ª edição. New York: McGraw-Hill.

Froberg, Carl-Erik (1970). *Introduction to Numerical Analysis*. 2ª edição. New York: Addison-Wesley.

Golub, G.H. e C.F. Van Loan (1996). *Matrix Computations*. 3ª edição. Baltimore: The Johns Hopkins University Press.

Hildebrand, F.B. (1956). *Introduction to Numerical Analysis*. 1ª edição. New York: McGraw-Hill.

Lapidus, Leon (1962). *Digital Computation for Chemical Engineers*. 1ª edição. New York: McGraw-Hill (ver página 5).

Massarani, Giulio (1970). *Introdução ao Cálculo Numérico*. 2ª edição. Rio de Janeiro: Ao Livro Técnico S.A. (ver página 5).

Nocedal, Jorge e Stephen J. Wright (2006). *Numerical Optimization*. 2ª edição. Berlin: Springer.

Press, William H. et al. (2007). *Numerical Recipes: The Art of Scientific Computing*. 3ª edição. Cambridge: Cambridge University Press.

Índice Remissivo