UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS

CÁLCULO NUMÉRICO PARA CONTROLE E AUTOMAÇÃO

Versão preliminar

Eduardo Camponogara
Eugênio de Bona Castelan Neto

Florianópolis, Agosto de 2008

Agradecimentos

Agradeço os acadêmicos Tiago Villaça Vianna Ferreira (Engenharia de Controle e Automação Industrial) e Maurício Rangel Guimarães Serra (Mestrado em Engenharia Elétrica) pelo auxílio na transcrição das notas de aula em LaTeX. Augusto Marasca de Conto (Mestrado em Engenharia Elétrica) realizou estágio de docência em Cálculo Numérico, deixando suas contribuições na parte de fatoração LU de matrizes.

Agradecimentos especiais vão para os monitores Felipe Lucci e Vinicius Strey que muito se empenharam na disciplina, contribuindo ao texto, auxiliando os alunos e não medindo esforços para facilitar a aprendizagem.

Sumário

Lista de Figuras

Lista de Tabelas

Notação

Capítulo 1

Introdução ao Estudo da Matemática Numérica

1.1 Natureza e Objetivos da Matemática Numérica

Antes do advento dos computadores eletrônicos, nos anos quarenta, métodos numéricos eram impraticáveis. Mesmo antes dos anos quarenta, dispositivos mecânicos tinham sido desenvolvidos para executar cálculos como, por exemplo, as máquinas desenvolvidas pela IBM para serem usadas no censo americano. No início dos anos quarenta, durante a Segunda Guerra Mundial, computadores eletrônicos foram desenvolvidos nos E.U.A., permitindo o cálculo automático de tabelas para artilharia naval e solução de outros problemas logísticos e militares. Nasce, portanto, o interesse científico e tecnológico por métodos numéricos.

A necessidade de um tratamento particular para os métodos numéricos se refere ao fato de que as propriedades básicas da aritmética real não valem mais quando executadas no computador. Números com infinito dígitos como, por exemplo, $\pi = 3.1415\ldots$, é representado por um número finito de dígitos na aritmética computacional—o computador tem precisão finita.

A matemática computacional é a área da matemática que se preocupa com o desenvolvimento, emprego e estudo de métodos numéricos, podendo ser subdivida em:

Matemática Computacional: estudo da matemática do ponto de vista computacional.

Matemática Numérica: parte da matemática computacional que se preocupa com o desenvolvimento de algoritmos para resolução aproximada

de problemas, a qual utiliza como sistema de operações o conjunto $\{+, -, /, *\}$ de operadores matemáticos.

Matemática Simbólica: busca a solução analítica de problemas matemáticos, por exemplo, a solução analítica da integral:

$$\int x^2 \cdot dx = \frac{x^3}{3}$$

Matemática Gráfica: trabalha com modelos gráficos buscando solução na forma gráfica.

Matemática Intervalar: trata dados na forma de intervalos, buscando controlar os limites de erro da matemática numérica.

Nesta disciplina nos concentramos na matemática numérica, onde estudamos processos numéricos para a resolução de problemas visando a máxima economia e confiabilidade em termos de fatores envolvidos, tais como:

- tempo de execução
- memória utilizada
- erros de arredondamento

1.2 Algoritmos

Informalmente, um algoritmo é um procedimento computacional bem definido que toma um valor (ou seqüência de valores) como entrada e produz um valor (ou seqüência de valores) como saída. Um algoritmo pode ser visto como uma ferramenta para resolver um problema computacional bem-definido, ou seja, um problema cuja especificação nos dá em termos gerais a relação desejada entre entrada e saída. Se espera que os passos de um algoritmo sejam executados no máximo um número finito de vezes, mas por questões práticas, se deseja que este número seja o menor possível e bem-comportado.

1.2.1 O Problema da Ordenação

Ilustraremos aqui as noções de problema computacional e algoritmo através do problema de ordenação.

Entrada ou Instância do Problema: uma seqüência de n números $(a_1, a_2, \ldots, a_n)$

Saída: a permutação (reordenação) $(a'_1, a'_2, \ldots, a'_n)$ da entrada tal que: $a'_1 \leqslant a'_2 \leqslant \ldots \leqslant a'_n$.

Dada uma sequência de entrada $(31, 41, 59, 26, 41, 58)$ um algoritmo de ordenação produz a saída $(26, 31, 41, 41, 58, 59)$.

Um algoritmo é dito *correto* se, para toda a instância, o algoritmo termina com a saída correta.

Algoritmo de Ordenação

$$(a_1, a_2, \ldots, a_n) \rightarrow \boxed{\text{SORT}} \rightarrow (a'_1, a'_2, \ldots, a'_n)$$

A implementação do algoritmo de ordenação pode ser definida como:

Insertion_Sort(A)

1: **for** $j \leftarrow 2$ to $length[A]$ **do**
2: $key \leftarrow A[j]$
3: $i \leftarrow j - 1$
4: **while** $i > 0$ & $A[i] > key$ **do**
5: $A[i + 1] \leftarrow A[i]$
6: $i \leftarrow i - 1$
7: **end while**
8: $A[i + 1] \leftarrow key$
9: **end for**

Abaixo, a demonstração de seu funcionamento, considerando a entrada (5,2,4,6,1,3):

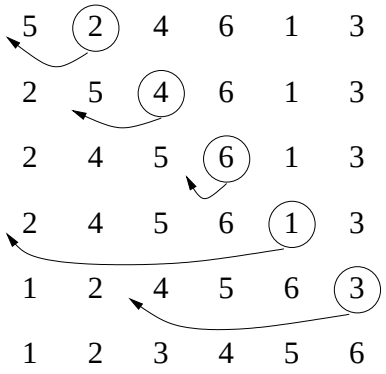


Figura 1.1: Algoritmo de Ordenação

Análise do Algoritmo de Ordenação: Tempo de Execução

O tempo de execução de um algoritmo (*"running time"*) para uma entrada particular consiste do número de operações primitivas ou passos executados. O tempo de execução do algoritmo de ordenação depende da entrada:

- A ordenação de 1000 números leva mais tempo do que a ordenação de 10 números.
- O tempo de ordenação de duas entradas pode variar de acordo com o grau de ordenação das entradas.

Para calcular o tempo de execução $T(n)$ para uma entrada de tamanho n, somamos todas as partes (linhas de 1 a 7), onde t_j é o número de vezes que o laço **while** é executado para aquele valor j.

Insertion_Sort(A)

Linha	Código	Custo	Repetições
1	for $j \leftarrow 2$ to $length[A]$	c_1	n
2	$key \leftarrow A[j]$	c_2	$n-1$
3	$i \leftarrow j-1$	c_3	$n-1$
4	while $i > 0$ & $A[i] > A[j]$	c_4	$\sum_{j=2}^{n} t_j$
5	$A[i+1] \leftarrow A[i]$	c_5	$\sum_{j=2}^{n} (t_j - 1)$
6	$i \leftarrow i-1$	c_6	$\sum_{j=2}^{n} (t_j - 1)$
7	$A[i+1] \leftarrow key$	c_7	$n-1$

Quanto tempo leva para ordenar as seqüencias abaixo?

$$s_1 = (1, 2, 3, 4, 5, 6)$$
$$s_2 = (6, 5, 4, 3, 2, 1)$$

Melhor Caso: percebe-se que a entrada s_1 já está ordenada portanto $t_j = 1$, caracterizando o melhor caso. O tempo de execução no melhor caso pode ser expresso por:

$$T(n) = an + b$$

Pior Caso: já a entrada s_2 está no maior grau de desordem possível para o tipo de entrada em questão sendo $t_j = j$, o pior caso. O tempo de execução no pior caso pode ser expresso por:

$$T(n) = an^2 + bn + c$$

Caso Médio: há também o caso médio, que pode ser encontrado sendo a média em uma distribuição das probabilidades de entrada.

Análise do Algoritmo de Ordenação: Ordem de Crescimento

A ordem de crescimento pode ser vista como a taxa de crescimento do tempo de execução em relação ao tamanho da entrada. Logo, consideramos o termo dominante da fórmula $T(n) = an^2 + bn + c$, uma vez que os termos de menor ordem são insignificantes para entradas grandes. Dizemos que o tempo de execução é $\Theta(n^2)$.

1.3 Algoritmos Numéricos

Da mesma forma que a solução numérica de equações lineares é fundamental à solução de equações não-lineares, algoritmos numéricos são fundamentais à solução de diversos problemas encontrados em engenharia, como a identificação de sistemas, tratamento de sinais e simulação. Abaixo seguem as caracteríticas desejadas dos algoritmos.

1.3.1 Inexistência do Erro Lógico

Um algoritmo não apresenta erro lógico se este sempre produz o resultado correto. Considere o exemplo abaixo.

Problema: procura-se a solução x^* de $ax = b$

Algoritmo ingênuo: $x^* = \frac{b}{a}$

Algoritmo correto:

Equação-Linear(a, b)

1: **if** $a = 0$ **then**
2: **if** $b = 0$ **then**
3: Imprima "identidade"
4: **else**
5: Imprima "contradição"
6: **end if**
7: **else**
8: Retorne $x^* = \frac{b}{a}$
9: **end if**

1.3.2 Inexistência do Erro Operacional

O algoritmo pode falhar por violar restrições físicas da máquina. No que segue desenvolvemos um exemplo ilustrativo. Seja T o conjunto de números possíveis de serem representados por uma máquina onde:

a) $\forall x \in T, -x \in T$

b) $t_1 = \inf\{x : x \in T \wedge x > 0\}$

c) $t_2 = \sup\{x : x \in T \wedge x > 0\}$

Se temos valores y tais que $|y| < t_1$ dizemos que ocorreu "underflow" ou se $|y| > t_2$ dizemos que ocorreu "overflow".

Considere o problema computacional no qual procuramos $|z| = \sqrt{x^2 + y^2}$. Se implementarmos diretamente a fórmula acima dependendo dos valores x ou y, podemos ter overflow em x^2 ou y^2, embora valha $\sqrt{x^2 + y^2} < t_2$.

O algoritmo abaixo procura contornar estes problemas de "overflow".

Norma-Vetorial(x, y)
1: **if** $x = y = 0$ **then**
2: $z = 0$
3: **else**
4: **if** $|x| \geqslant |y|$ **then**
5: $z = |x|\sqrt{1 + \left(\frac{y}{x}\right)^2}$
6: **else**
7: $z = |y|\sqrt{1 + \left(\frac{x}{y}\right)^2}$
8: **end if**
9: **end if**
10: Retorne z

1.3.3 Quantidade Finita de Cálculos

Em algoritmos iterativos, é necessário que se estabeleça um critério de parada e se prove convergência. Um algoritmo não pode executar indefinidamente e quando ele pára se espera que este tenha produzido o resultado esperado.

Considere o problema de determinar, pelo método de Newton, uma raiz da equação $f(x) = sign(x) \cdot \sqrt{\|x\|} = 0$, onde:

$$\begin{cases} sign(x) = 1 & \text{se} \quad x > 0 \\ sign(x) = 0 & \text{se} \quad x = 0 \\ sign(x) = -1 & \text{se} \quad x < 0 \end{cases}$$

Um algoritmo problemático é dado por:

Newton(f, x_0, γ)

1: $j \leftarrow 0$
2: **while** $|f(x_j)| > \gamma$ **do**
3: **if** $f'(x_j) \neq 0$ **then**
4: $x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)}$
5: **else**
6: Pare e imprima "indefinição"
7: **end if**
8: $j \leftarrow j + 1$
9: **end while**
10: Retorne $\{j, x_j\}$

O gráfico da função e iterandos consecutivos gerados pelo algoritmo acima são ilustrados na Figura 1.2. A seqüência de iterandos gerados a partir de $x_0 = 0.8$ é dada na Figura 1.3.

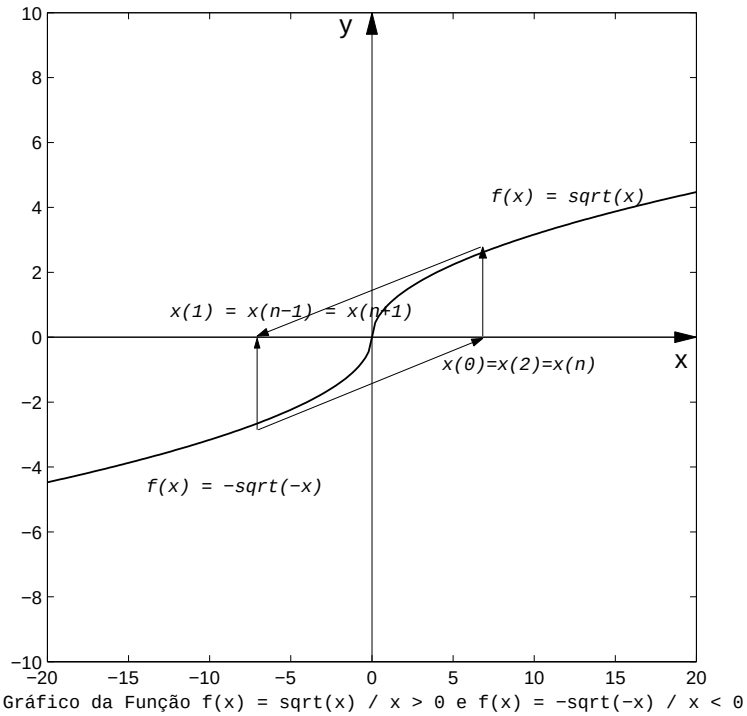


Figura 1.2: Exemplo de função para a qual o algoritmo de Newton pode iterar indefinidamente

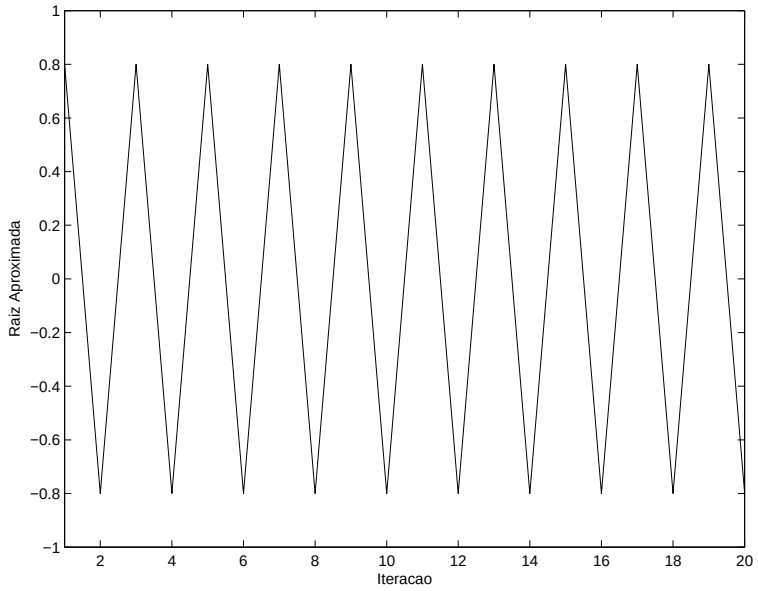


Figura 1.3: Seqüência de iterandos gerados pelo algoritmo de Newton.

1.3.4 Existência de um Critério de Exatidão

É fundamental que o algoritmo forneça, de antemão, um critério de exatidão em função das limitações de precisão das máquinas. Ou seja, se deseja que o algoritmo forneça:

$$\text{Resultado} = \text{Valor Aproximado} + \text{Erro}$$

1.3.5 Independência da Máquina

É desejável que o algoritmo produza o mesmo resultado quando executado em diferentes máquinas. A constante de Euler $e = 2.718281828\ldots$, por exemplo, terá representação distinta em diferentes máquinas. Assim, não se deve utilizar o valor, mas sim a representação $e = \exp(1)$ que corresponde ao valor adotado pelo compilador.

1.3.6 Com Precisão Infinita, os Limites de Erro Devem Convergir para Zero

Esta exigência estabelece a dependência entre a solução ideal em $\mathbb{R}$ e a solução de máquina. Considere o problema de determinar $\sin(\alpha) = x$ dado $\alpha \in \mathbb{R}$. Um algoritmo que não satisfaz a condição de erro nulo com precisão infinita é dado abaixo.

sin(α)
1: $x = 0 \pm 1$
2: Retorne $\{\alpha, x\}$

O algoritmo acima é insensível à precisão da máquina e, portanto, não satisfaz o critério desejado.

1.3.7 Eficiência

Quando se deseja encontrar a solução para um problema, sempre visamos obter economia de rescursos envolvidos. Alguns fatores relevantes são:

- tempo de execução;
- exatidão;
- volume de dados;
- dificuldade de representação; e
- eficácia.

Fazer contas com os dedos da mão, por exemplo, é eficaz mas não é eficiente para cálculos aritméticos não triviais.

Outro exemplo se refere ao algoritmo de Cramer para a solução de sistemas de equações lineares: $Ax = b$, com $A \in \mathbb{R}^{n \times n}$. Os passos do algoritmo são:

1) calcule o determinante Δ da matriz dos coeficientes;
2) calcule os n determinantes Δx_j resultantes da substituição da coluna j da matriz dos coeficientes pelo vetor b; e
3) a solução $x = (x_1, x_2, \ldots, x_n)$ é dada por $x_j = \frac{\Delta x_j}{\Delta}$, $j = 1, \ldots, n$.

O algoritmo de Cramer acima executará $(n+1)!(n+1)$ operações aritméticas mas, por outro lado, o algoritmo de Gauss termina após n^3 operações.

1.4 Passos para a Resolução de um Problema

Os passos recomendados para resolução de um problema através de métodos numéricos, ilustrados na Figura 1.4, são:

a) Modelo matemático do problema

b) Necessidade de simplificação

c) Uso de valores já conhecidos

d) Parâmetros de equações termodinâmicas

e) Escolha ou desenvolvimento do algoritmo

f) Definição de parâmetros do algoritmo

   - tolerância
   - solução inicial

g) Como a maioria dos problemas não tem solução direta e precisa, faz-se o uso de métodos iterativos.

   - critério de parada
   - número de iterações

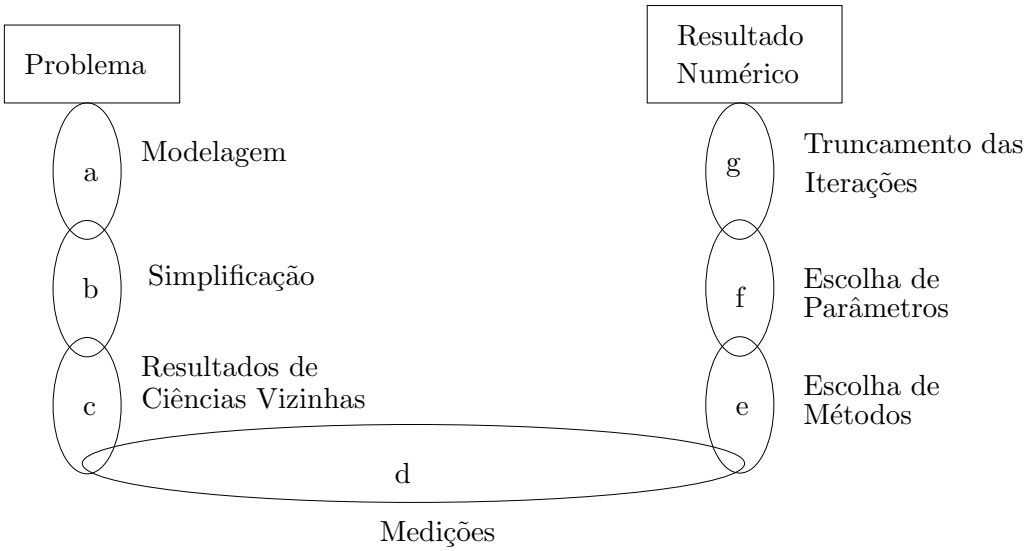


Figura 1.4: Passos para solução de um problema.

1.4.1 Referências

O texto deste capítulo é um sumário do Capítulo 1 do Texto de Cláudio e Marins [1]. O leitor pode consultar este texto para uma discussão mais ampla sobre Matemática Numérica.

Capítulo 2

Introdução à Aritmética de Máquina

Neste capítulo trataremos da representação aproximada de números reais em notação digital nas máquinas computacionais. Serão definidos sistema de ponto flutuante, funções de arredondamento e tipos de erros que surgem com arredondamentos. Discutiremos a diferença entre precisão e exatidão. Também serão apresentadas questões relativas à instabilidade de algoritmos e de problemas.

2.1 O Sistema de Ponto Flutuante

Números reais são aproximados por números racionais em máquinas digitais de precisão finita (número de dígitos limitados), incorrendo erros que podem ser amplificados à medida que operações aritméticas são executadas. Devemos, portanto, entender como números são representados no computador e a maneira com que as operações são executadas. Tal conhecimento servirá de base para análise e depuração de algoritmos, bem como validação de resultados obtidos através de métodos numéricos.

Exemplo:

Vamos utilizar várias máquinas para calcular a expressão abaixo:

$$\begin{aligned} H &= 1/2 \\ X &= 2/3 - H \\ Y &= 3/5 - H \\ E &= (X + X + X) - H \\ F &= (Y + Y + Y + Y + Y) - H \\ G &= F/E \end{aligned} \tag{2.1}$$

Os resultados destas operações nas máquinas HP-25, SR-50, PCRII, IBM-4341, e Matlab (IBM-PC) aparecem na Tabelas 2.1 e 2.2. Podemos observar que os resultados variam significativamente, dependendo da capacidade de representação destas máquinas.

Tabela 2.1: Operações Aritméticas

HP 25	SR50	PCRII
$H = 0.5$	$H = 0.5$	$H = 0.5$
$X = 0.166666667$	$X = 0.166666667$	$X = 0.166666667$
$Y = 0.1$	$Y = 0.1$	$Y = 0.1$
$E = 10^{-10}$	$E = 2 \cdot 10^{-13}$	$E = 10^{-10}$
$F = 0$	$F = 0$	$F = 0$
$G = nd$	$G = nd$	$G = nd$

Tabela 2.2: Operações Aritméticas

IBM4341	Matlab
$H = 0.5$	$H = 0.5$
$X = 0.16666666$	X = 0.16666666666667
$Y = 0.1$	$Y = 0.1$
$E = -0.119209 \cdot 10^{-6}$	$E = -1.110223024625157 \cdot 10^{-16}$
$F = -0.178813 \cdot 10^{-6}$	$F = -1.110223024625157 \cdot 10^{-16}$
$G = 0.6666\ldots$	$G = 1$

Abaixo são dadas definições fundamentais e a especificação de sistema de ponto flutuante.

Definição 2.1 *Um $x \in \mathbb{R}$ é dito número de ponto flutuante normalizado se valerem:*

1) $x = m \cdot b^e$

2) $m = \pm 0 \cdot d_1 d_2 \ldots d_n, n \in \mathbb{N}$

3) $1 \leqslant d_1 \leqslant b-1$ *e* $0 \leqslant d_j \leqslant b-1$, $j = 2, \ldots, n$

4) $e_1 \leqslant e \leqslant e_2$, *sendo* $e_1 \leqslant 0, e_2 \geq 1, e_1, e_2 \in \mathbb{Z}$

onde:

- b *é chamado de base,* $b \geq 2$
- e *é chamado de expoente,* e_1 *é o menor expoente e* e_2 *é o maior expoente*
- m *é chamado de mantissa*
- n *é o número máximo de dígitos usados na representação do número*
- d_j, $j = 1, \ldots, n$, *são os dígitos do número*

Definição 2.2 *A união de todos os números de ponto flutuante com o ZERO, que é representado na seguinte forma:* $0 = 0.0000\ldots0 \cdot b^{e_1}$ *é chamado de sistema de ponto flutuante. Usualmente, procuramos representar um sistema de ponto flutuante por* $F = F(b, n, e_1, e_2)$, *onde* e_1 *e* e_2 *são respectivamente o menor e o maior expoente,* b *é a base e* n *é a precisão.*

Alguns exemplos de sistemas de ponto flutuante:

1) HP25, $F(10, 9, -98, 100)$
2) IBM 360/370, $F(16, 6, -64, 63)$
3) B6700, $F(8, 13, -51, 77)$

Algumas propriedades de sistemas de ponto flutuante são:

Menor número em módulo: $0.1 \cdot b^{e_1}$

Maior número: $0.[b-1][b-1]\ldots[b-1] \cdot b^{e_2}$

Cardinalidade de $F = F(b, n, e_1, e_2)$: $\#F = 2(b-1)(b^{n-1})(e_2 - e_1 + 1) + 1$, que pode ser obtido adicionando-se as parcelas:

- O número de mantissas positivas é dado por $(b-1)(b^{n-1})$

Tabela 2.3: Elementos do sistema de ponto flutuante $F(2, 3, -1, 2)$

		mantissa m			
e	b^e	0.100	0.101	0.110	0.111
-1	1/2	1/4	5/16	3/8	7/16
0	1	1/2	5/8	3/4	7/8
1	2	1	5/4	3/2	7/4
2	4	2	5/2	3	7/2

- Como cada uma dessas mantissas pode ter um dos $(e_2 - e_1 + 1)$ expoentes possíveis, temos ao todo $(b-1)(b^{n-1})(e_2 - e_1 + 1)$ números possíveis
- Logo, incluindo os negativos e o zero, obtemos $\#F = 2(b-1)(b^{n-1})(e_2 - e_1 + 1) + 1$

Para qualquer mantissa m, vale $b^{-1} \leqslant |m| < 1$, pois:

- $|m| < 1$, pois toda a mantissa tem como primeiro dígito o ZERO
- $|m| \geq b^{-1}$, pois se $|m| < b^{-1}$, não teríamos um número normalizado, pois o primeiro dígito após o ponto não é nulo.

Exemplo:

Seja $F = F(2, 3, -1, 2)$. Para este sistema, as mantissas são: 0.100, 0.101, 0.110, e 0.111. Por outro lado, os expoentes admissíveis são -1, 0, 1 e 2. Assim temos os seguintes números positivos:

- $(0.100 \times 2^{-1})_2 = (0.01)_2 = 0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} = 1/4$
- $(0.100 \times 2^{0})_2 = (0.1)_2 = 0 \times 2^0 + 1 \times 2^{-1} = 1/2$
- $(0.100 \times 2^{1})_2 = (1)_2 = 1$
- $(0.100 \times 2^{2})_2 = (10)_2 = 2^1 = 2$
- e assim sucessivamente, obtendo a Tabela 2.3 que também são apresentados na Figura 2.1.

Outras noções importantes se referem aos limites de representação de um sistema de ponto flutuante.

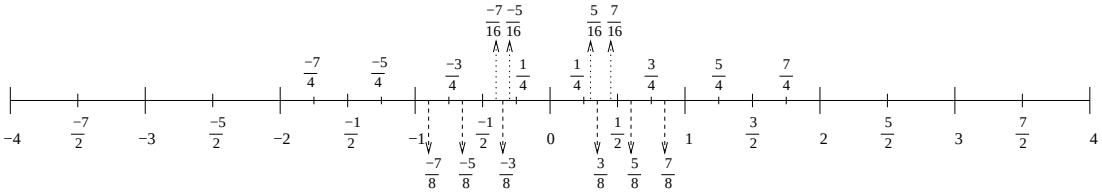


Figura 2.1: Elementos do sistema de ponto flutuante $F = F(2, 3, -1, 2)$.

Região de Underflow: região situada entre o maior número de ponto flutuante negativo e o ZERO e, simetricamente, entre o menor número de ponto flutuante positivo e o ZERO.

Região de Overflow: regiões situadas aquém do menor número de ponto flutuante negativo e além do maior número de ponto flutuante positivo.

Exemplo:

Seja $F = F(2, 3, -1, 2)$ um sistema de ponto flutuante. Tomemos em F, $x = \frac{5}{4}$ e $y = \frac{3}{8}$. Note que $z = x + y = \frac{5}{4} + \frac{3}{8} = \frac{13}{8}$ é tal que $z \notin F$, pois $\frac{13}{8} = (0.1101 \times 2^1)_2$ que possui um dígito a mais na mantissa do que o permitido. Na realidade, podemos escolher entre $\frac{3}{2} = (0.110 \times 2^1)_2$ ou $\frac{7}{4} = (0.111 \times 2^1)_2$, o que dá origem a diferentes tipos de arredondamentos—erros são cometidos ao se aproximar o número $z = x + y$ com um elemento de F.

Notação: Seja $\{\oplus, \otimes, \ominus, \oslash\}$ o conjunto de operações executados por um algoritmo de ponto flutuante equivalentes às operações do conjunto $\{+, -, /, \times\}$. Podemos verificar facilmente que: $x \oplus y \neq x + y$ e $x \otimes y \neq x \times y$.

Considere o sistema $F = F(2, 5, -98, 100)$ e os números:

$$\begin{aligned} (0.1)_{10} &= (0.0001110110011\ldots)_2 \notin F \\ (0.1)_{10} &\simeq (0.11101 \times 2^{-3})_2 \in F. \end{aligned}$$

Somando (0.11101×2^{-3}) sucessivamente dez vezes, teremos $(0.11111)_2 = (0.96875)_{10} \neq (1.0)_{10}$.

Exemplo:

Para um sistema de ponto flutuante $F = F(2, 3, -1, 2)$, seja:

$$x = \frac{5}{8}, \quad y = \frac{3}{8}, \quad e \quad z = \frac{3}{4},$$

então:

$$\begin{aligned}(x \oplus y) \oplus z &= ((0.101 \times 2^0) \oplus (0.110 \times 2^{-1})) \oplus (0.110 \times 2^0) \\ &= (0.101 \oplus 0.011) \oplus 0.110 \\ &= 1.0 \oplus 0.110 = 1.11\end{aligned}$$

$$\begin{aligned}x \oplus (y \oplus z) &= 0.101 \oplus (0.011 \oplus 0.110) \\ &= 0.101 \oplus 1.001 \\ &= 1.101 \\ &= 1.10\end{aligned}$$

Logo $(x \oplus y) \oplus z = x \oplus (y \oplus z)$

Exemplo:

Para o sistema de ponto flutuante $F = F(2, 3, -1, 2)$, seja

$$x = \frac{7}{8}, \quad y = \frac{5}{4}, \quad e \quad z = \frac{3}{8},$$

então, podemos verificar que:

$$\begin{aligned}x \otimes (y \oplus z) &= 0.111 \otimes (1.01 \oplus 0.011) \\ &= 0.111 \otimes 1.101 \\ &= 0.111 \otimes 1.10 \\ &= 0.111 \oplus 0.0111 \\ &= 1.0101 \\ &= 1.01 \\ (x \otimes y) \oplus (x \otimes z) &= (0.111 \otimes 1.01) \oplus (0.111 \otimes 0.011) \\ &= 1.00011 \oplus 0.010101 \\ &= 1.00011 \oplus 0.010101 \\ &= 1.00 \oplus 0.0101 \\ &= 1.0101 \\ &= 1.01\end{aligned}$$

Contudo, neste caso $x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z)$.

2.1.1 Exercícios

i. Qual é o valor exato da expressão (2.1)?

ii. Considere os sistemas de ponto flutuante abaixo:

   (a) HP25, $F(10, 9, -98, 100)$;

   (b) IBM 360/370, $F(16, 6, -64, 63)$; e

   (c) B6700, $F(8, 13, -51, 77)$.

   Questões:

   (a) Qual é cardinalidade de cada um destes sistemas?

   (b) Qual é o menor número em módulo que pode ser representado?

   (c) Qual é o maior número em módulo?

   (d) Qual é a região de *overflow*?

   (e) Qual é a região de *underflow*?

iii. Em um sistema de ponto flutuante $F = F(2, 3, -1, 2)$, seja $x = 0.110 \times 2^1$, $y = 0.101 \times 2^{-1}$ e $z = 0.100 \times 2^1$. Calcule $(x \oplus y) \ominus z$ e $(x \ominus z) \oplus y$, utilizando o sistema F e truncamento.

iv. Para o sistema de ponto flutuante $F = F(2, 3, -1, 2)$, encontre $x, y, z \in F$ tal que $x \otimes (y \oplus z) \neq (x \otimes y) \oplus (x \otimes z)$.

2.2 Arredondamentos

Conforme visto na discussão acima, há diferentes maneiras de se aproximar um número real para um número de ponto flutuante. Surge a questão de como se realizar tal aproximação.

Definição 2.3 *Seja $F(b, n, e_1, e_2)$ um sistema de ponto flutuante. Uma função $\square : \mathbb{R} \to F$ é considerada um arredondamento se valer: $\forall x \in F, \square(x) = x$*

Tipos de arredondamento

- Arredondamento para cima ou por excesso: Δx
- Arredondamento para baixo ou por falta: ∇x
- Arredondamento para o número de máquina mais próximo: ox

Exemplo

Seja $F = F(2, 3, -1, 2)$ o sistema de ponto flutuante. O número $\frac{9}{8} \notin F$, pois:

$$\frac{9}{8} = (1.125)_{10} = (0.1001 \times 2^1)_2.$$

Podemos arredondar $\frac{9}{8}$ para $(0.100 \times 2^1)_2 = (1.0)_{10}$ ou para $(0.101 \times 2^1)_2 = (\frac{5}{4})_{10} = (1.25)_{10}$. No primeiro caso, temos $\nabla(\frac{9}{8}) = (0.100 \times 2^1)$, já no segundo caso temos $\Delta(\frac{9}{8}) = (0.101 \times 2^1)$

Exemplo:

Seja $F = F(10, 4, -98, 10)$ o sistema de ponto flutuante, e sejam:

$$\begin{aligned} x &= 0.333333 \\ y &= 0.348436 \\ z &= 0.666666 \end{aligned}$$

Então obtemos os seguintes números para os diferentes arredondamento:

$$\begin{array}{rclrclrcl} \nabla(x) & = & 0.3333 & \nabla(y) & = & 0.3484 & \nabla(z) & = & 0.6666 \\ \Delta(x) & = & 0.3334 & \Delta(y) & = & 0.3485 & \Delta(z) & = & 0.6667 \\ ox & = & 0.3333 & oy & = & 0.3484 & oz & = & 0.6667 \end{array}$$

Definição 2.4 *Um arredondamento $\square : \mathbb{R} \to F$ é dito por falta se valer:* $\forall x \in \mathbb{R}, \square(x) \leqslant x$.

Definição 2.5 *Um arredondamento $\square : \mathbb{R} \to F$ é dito por excesso se valer:* $\forall x \in \mathbb{R}, \square(x) \geq x$.

Definição 2.6 *Um arredondamento é dito monotônico se valer:* $\forall x, y \in \mathbb{R}, x \leqslant y \Rightarrow \square(x) \leqslant \square(y)$.

Note que $\nabla(x)$ é um arredondamento monotônico por falta, enquanto $\Delta(x)$ é um arredondamento monotônico por excesso.

2.3 Erros

Toda vez que executamos um arredondamento que não admite uma representação exata em F, cometemos um erro. Há várias causas de erro. Aqui vamos estudar três tipos de erro:

Erros Inerentes: aparecem na criação ou simplificação de um modelo matemático de determinado sistema (erros na medição, identificação).

Erros de Discretização: erros cometidos quando se substitui qualquer processo infinito por um processo finito ou discreto como, por exemplo:

$$e = \sum_{i=0}^{\infty} \frac{1}{i!}$$

é aproximado com a série finita

$$e = \sum_{i=0}^{T} \frac{1}{i!}$$

Erros de Arredondamento: surgem quando trabalhamos com máquinas digitais para representar os números reais. Em geral trabalhamos com arredondamento para o número de ponto flutuante mais próximo ou com o arredondamento por falta.

A diferença entre o valor arredondado e o valor exato pode ser medida pelo erro absoluto ou pelo erro relativo, cujas definições são dadas a seguir.

Definição 2.7 *O erro absoluto E_A é dado por: $E_A = |\square(x) - x|$.*

Definição 2.8 *O erro relativo E_R é dado por $E_R = \frac{|\square(x)-x|}{|x|}$ ou $E_R = \frac{|\square(x)-x|}{|\square(x)|}$*

Erros relativos são mais usados que os erros absolutos. Um exemplo de erro absoluto e erro relativo é dado abaixo:

$$\begin{aligned} x &= 0.00006 \\ \square(x) &= 0.00005 \\ E_A &= 0.00001 \\ E_R &= \frac{0.00001}{0.00005} \\ &= 0.2 \end{aligned}$$

Alguns resultados teóricos podem ser estabelecidos sobre limites de erros levando em consideração propriedades do sistema de ponto flutuante.

Teorema 2.1 *Seja $F = F(b, n, e_1, e_2)$ um sistema de ponto flutuante. Então vale:*

$$\forall x \in \mathbb{R}, b^{e_1-1} \leqslant |x| \leqslant B^* \Rightarrow \frac{|\square(x) - x|}{|\square(x)|} \leqslant \mu,$$

onde: B^* *é o maior número em módulo do sistema de ponto flutuante* F*, e*

$$\mu = \begin{cases} \frac{1}{2}b^{1-n}, & \text{no caso de } \square x = ox \\ b^{1-n}, & \text{no caso de } \square x = \nabla x \end{cases}$$

O teorema acima só é válido quando x está dentro do espectro de representação de F, ou seja, $b^{e_1-1} \leqslant |x| \leqslant B^*$. No caso de underflow $|x| < b^{e_1-1}$ ou overflow $|x| > B^*$, não é possível de se obter cotas acima para erro relativo.

2.4 Dígitos Significativos Exatos

Na prática, quando obtemos um resultado de uma expressão numérica avaliada numa máquina e não podemos saber o valor exato, torna-se impossível calcular o erro relativo ou absoluto.

Definição 2.9 *Em um sistema decimal, um dígito é significativo se for* $1, 2, \ldots, 9$*. O dígito* 0 *é significante, exceto quando for usado para fixar a vírgula, ou o ponto decimal, ou preencher o lugar de dígitos descartados.*

Exemplo

$$\begin{aligned} 0.008735 &\rightarrow \text{4 dígitos significativos} \\ 30.357 &\rightarrow \text{5 dígitos significativos} \\ 23.000 &\rightarrow \text{2 dígitos significativos} \end{aligned}$$

Definição 2.10 *Um dígito significativo é exato se, arredondando-se o número aproximado para uma posição imediatamente após aquela posição do dígito, isso fizer com que o erro absoluto não seja maior do que a meia unidade naquela posição do dígito. Abreviamos o dígito significativo exato por DIGSE.*

Exemplo Os números 0.66667 e 0.666998 são aproximações para $\frac{2}{3}$. No entanto, todos os dígitos significativos do primeiro são exatos, enquanto no segundo só os três primeiros. Abaixo calculamos os dígitos significativos exatos de cada aproximação.

Primeiro Caso: 0.66667

$$\begin{array}{lrcll} \text{primeiro dígito} & |0.66 - 0.6666\ldots| & = & |-0.006666\ldots| & < 0.05 \\ \text{segundo dígito} & |0.666 - 0.6666\ldots| & = & |-0.0006666\ldots| & < 0.005 \\ \text{terceiro dígito} & |0.6666 - 0.6666666\ldots| & = & |-0.00006666\ldots| & < 0.0005 \\ \vdots & & & & \\ \text{quinto dígito} & |0.666670 - 0.666666\ldots| & = & |+0.00000333\ldots| & < 0.000005 \end{array}$$

Logo todos os dígitos são significativos exatos.

Segundo Caso: 0.666998 Para o primeiro dígito 9 temos

$$|0.66699 - 0.66666\ldots| = |0.000323\ldots| \not\leqslant 0.00005$$

logo, o primeiro dígito 9 já não é exato.

Teorema 2.2 *Se $E_R \leqslant \frac{1}{2}b^{-m}$, então o número é correto em m dígitos significativos exatos.*

2.4.1 Exercícios

i. No sistema $F(2, 3, -1, 2)$, represente $a = 5/7$ e $b = 2/3$, utilizando os arredondamentos $\nabla(x)$, $\Delta(x)$ e $o(x)$.

ii. Para o item anterior, calcule o erro absoluto e relativo cometido ao se aproximar a com $\square(a)$, e b com $\square(b)$, para cada função de arredondamento.

iii. Calcule o número de dígitos significativos exatos das aproximações $x = 3.14169$ e $y = 3.141412$ para π.

iv. Podemos afirmar que o arredondamento $o(x)$ é monotônico?

2.5 Precisão e Exatidão de Máquinas Digitais

Conforme vimos, para cada máquina, calculadora ou computador há um sistema de ponto flutuante associado. Este sistema automaticamente define a precisão da máquina.

Definição 2.11 *(Épsilon de máquina) O* épsilon *da máquina é o menor número de ponto flutuante tal que:*

$$1 + \epsilon > 1.$$

Definição 2.12 *A precisão de uma máquina digital é definida como o número de dígitos da mantissa dessa máquina.*

Definição 2.13 *Exatidão é uma medida de perfeição do resultado.*

Tabela 2.4: Aproximações do número π

Aproximação x_j	$DIGSE(x_j, \pi)$
3.1410	3.4
3.1411	3.5
3.1412	3.6
3.1413	3.7
3.1414	3.9
3.1415	4.2
3.1416	5.3
3.1417	4.2
3.1418	3.7
3.1419	3.6

A exatidão de um resultado depende da precisão da máquina e do método utilizado para a obtenção do resultado. Quando não conhecemos o valor exato temos, em geral, que se $x = \lim_{j\to\infty} x_j$, então o número de dígitos significativos de x_j em relação a x_{j+1} é dado por:

$$DIGSE(x_j, x_{j+1}) = -\left[0.3 + \log(\mu + \frac{|x_{j+1} - x_j|}{|x_j|})\right]$$

onde μ é a unidade de erro de arredondamento e $b = 10$ é a base.

Exemplo

Considere as aproximações para o número π, conforme Tabela 2.4.

Embora todas as aproximações possuam uma precisão de cinco dígitos, somente uma delas possue cinco dígitos significantes exatos. Logo, exatidão de um processo depende além da máquina, também do algoritmo.

Exemplo:

Seja o número irracional $\sqrt{2} = 1.414213562\ldots$.

a) 1.4142 é mais preciso e mais exato que 1.41, pois o primeiro tem maior número de casas decimais e aproxima melhor $\sqrt{2}$;

b) 1.4149 é mais preciso que 1.414, pois tem mais casas decimais, porém, é menos exato do que 1.414, já que o dígito 9 do primeiro não é exato.

Exercícios

i. Podemos afirmar que um resultado exato também é preciso? Podemos afirmar o oposto?

ii. Reproduza os valores da Tabela 2.4 utilizando uma plataforma para computação numérica como, por exemplo, Octave ou Matlab.

2.6 Instabilidade

Veremos agora uma série de problemas cujos diferentes modos de solução podem acarretar diferentes resultados. Quando os resultados obtidos não são aceitáveis, os erros podem ser causados:

a) pelos modelos ou entrada de dados (erros inerentes)

b) pelo arredondamento ou truncamento.

Instabilidade pode ser entendida como uma sensibilidade a perturbações e pode ocorrer tanto no problema em si como no algoritmo, isto é, na maneira de resolvê-lo.

2.6.1 Instabilidade dos Algoritmos

Exemplo 1: função e^x

A instabilidade de algoritmos pode ser ilustrada através do exemplo de se calcular a constante de Euler. Vamos calcular e e $e^{-5.5}$ pela série de Taylor. Dado que:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \tag{2.2}$$

Então para $x = 1$, temos:

$$e \cong 1 + 1 + 0.5 + \dots = 2.7183$$

Comparado a soma acima com o valor 2.718 281 828 obtido em uma calculadora, verificamos um erro relativo de 6.6×10^{-6}, que é bem pequeno.

Por outro lado, para $x = -5.5$, temos:

$$e^{-5.5} \cong 1 - 5.5 + 15.125 - 27.730 + \dots = 0.002\ 636\ 3.$$

Comparando agora com $e^{-5.5}$ dado por uma calculadora, temos que:

$$e^{-5.5} = 0.004\ 086\ 714\ 39,$$

portanto, o erro relativo é 0.35 que é bem maior que o erro anterior.

Qual a causa da diferença?

A causa do erro é uma combinação de dois fatores:

- somas de grandezas de diferentes ordens; e
- subtração de grandezas quase iguais.

Tal fenômeno é dito *cancelamento subtrativo* ou *cancelamento catastrófico*, que é bastante comum em cálculos. O cancelamento subtrativo não é a real causa do erro final da soma, ele apenas aumentou o efeito do erro final. Note que na primeira soma (para e), não houve tal aumento. Se mudarmos o cálculo de $e^{-5.5}$ para $\frac{1}{e^{5.5}}$ e utilizarmos as mesmas parcelas, obteremos 0.0040865 com erro relativo de 6.6×10^{-5}. Logo podemos utilizar a série de Taylor para argumentos positivos. Na prática também devemos utilizar um critério de parada mais cuidadoso do que o simples número de termos da série.

Qualidade da Aproximação

Seja $f(x) = \sum_{j=0}^{T} \frac{x^j}{j!}$ a aproximação de e^x obtida com a expansão de Taylor de ordem T em torno do ponto $x = 0$. Na Tabela 2.5 ilustramos os valores obtidos com $f(x)$ para $x \geqslant 0$ e, também, as aproximações obtidas para e^x com $f(x)$ e $1/f(-x)$ quando $x < 0$. Na mesma tabela são apresentados os valores de e^x computados através do Matlab, seguidos dos erros relativos induzidos por $f(x)$ e $1/f(-x)$ (apenas quando $x < 0$) em relação a e^x (Matlab). A partir dos valores apresentados na tabela podemos perceber que o erro resultante da aproximação $f(x)$ é elevado para $x < 0$.

Exemplo 2

Outro exemplo de instabilidade de algoritmo surge do cálculo da integral definida $l_n = \int_0^1 x^n e^{x-1} dx$ para $n = 1, 2, \ldots$ Integrando por partes, temos:

$$\begin{aligned} l_n &= \int_0^1 x^n e^{x-1} dx; \quad u = x^n \quad e \quad dv = e^{x-1} dx \\ &= [uv]_0^1 - \int_0^1 v du \\ &= [x^n e^{x-1}]_0^1 - \int_0^1 e^{x-1} n x^{n-1} dx \\ &= 1 - n l_{n-1}, \quad n = 2, 3, \ldots \end{aligned} \tag{2.3}$$

Tabela 2.5: Aproximações de e^x obtidas com $f(x)$ para $x \geqslant 0$ e com $f(x)$ e $1/f(-x)$ para $x < 0$.

x	$f(x)$	$1/f(-x)$	e^x (Matlab)	$\lvert e^x - f(x)\rvert/e^x$	$\lvert e^x - 1/f(-x)\rvert/e^x$
-40	50.810	4.2484e-018	4.2484e-018	10e020	5.4400e-016
-20	4.1736e-009	2.0612e-009	2.0612e-009	102.48	2.0066e-016
-10	4.5400e-005	4.5400e-005	4.5400e-005	7.2342e-007	2.9851e-016
-5	6.7379e-003	6.7379e-003	6.7379e-003	2.1369e-011	2.5746e-016
-2	0.13534	0.13534	0.13534	4.1018e-014	2.0509e-016
-1	0.36788	0.36788	0.36788	3.0179e-014	1.5089e-016
0	1.0000		1.0000	0.0000	
1	2.7183		2.7183	0.0000	
2	7.3891		7.3891	2.4040e-014	
5	1.4841e+002		1.4841e+002	1.9150e-014	
10	2.2026e+004		2.2026e+004	3.3033e-014	
20	4.8517e+008		4.8517e+008	2.4571e-014	

Podemos calcular analiticamente o valor de l_1, obtendo

$$\begin{aligned} l_1 &= \int_0^1 x e^{x-1} dx \\ &= [x e^{x-1}]_0^1 - \int_0^1 e^{x-1} dx \\ &= 1 \cdot e^0 - 0 \cdot e^{-1} - [e^{x-1}]_0^1 \\ &= 1 - e^0 + e^{-1} \\ &= 1/e. \end{aligned}$$

Usando $F = F(10, 6, -98, 99)$, temos os seguintes valores:

$$\begin{array}{lll} l_1 \simeq 0.367\,879 & & l_6 \simeq 0.127\,120 \\ l_2 \simeq 0.264\,242 & & l_7 \simeq 0.110\,160 \\ l_3 \simeq 0.207\,274 & & l_8 \simeq 0.118\,720 \\ l_4 \simeq 0.170\,904 & & l_9 \simeq -0.068\,480 \\ l_5 \simeq 0.145\,480 & & \end{array}$$

Olhando o integrando $x^9 e^{x-1}$, verificamos que é sempre positivo em $[0, 1]$ e, no entanto, o valor computado para l_9 foi *negativo*.

O que causou o erro?

Notemos que só foi feito um erro de arredondamento em l_1 quando $\frac{1}{e}$ foi tomado por 0.367879 em vez de 0.367879442. Como a fórmula está correta, o erro final é devido apenas a este erro cometido em l_1.

Observemos como tal erro ocorreu. Em l_2 o erro foi multiplicado por -2, depois em l_3 foi multiplicado por -3, etc. Então o erro de l_9 é exatamente $(-2)(-3)\ldots(-9) = 9!$. Sendo

$$E_1 = [\frac{1}{e} - 0.367879] = 4.412 \times 10^{-7},$$

teremos no final

$$4.412 \times 10^{-7} \cdot 9! = 0.1601$$

Logo, embora a fórmula esteja correta, ela é instável. Com relação ao acúmulo de erros, o algoritmo gerado é de má exatidão. Um algoritmo estável é dado por:

$$l_{n-1} = \frac{1 - l_n}{n}, \quad n = \ldots, 4, 3, 2. \tag{2.4}$$

Nesta fórmula, a cada passo, o valor do erro em l_n é decrescido por $\frac{1}{n}$ (em vez de multiplicado por n). Se começarmos com $n \gg 1$ voltaremos e então o erro inicial ou erros de arredondamento diminuirão a cada passo. Resta-nos saber qual será o valor inicial para l_n. Observamos que:

$$l_n = \int_0^1 x^n e^{x-1}.dx \leqslant \int_0^1 x^n.dx = \left[\frac{x^{n+1}}{n+1}\right]_0^1 = \frac{1}{n+1}.$$

Portanto, l_n tende a zero quando n tende ao infinito. Se aproximarmos l_{20} para zero e o usarmos como valor inicial, teremos:

$$\begin{array}{rcl@{\qquad}rcl}
l_{20} &\simeq& 0 & l_{14} &\simeq& 0.062\ 732\ 2 \\
l_{19} &\simeq& 0.050\ 000\ 0 & l_{13} &\simeq& 0.066\ 947\ 7 \\
l_{18} &\simeq& 0.050\ 000\ 0 & l_{12} &\simeq& 0.071\ 773\ 3 \\
l_{17} &\simeq& 0.052\ 777\ 8 & l_{11} &\simeq& 0.077\ 352\ 3 \\
l_{16} &\simeq& 0.055\ 719\ 0 & l_{10} &\simeq& 0.083\ 877\ 1 \\
l_{15} &\simeq& 0.059\ 017\ 6 & l_9 &\simeq& 0.091\ 612\ 3.
\end{array}$$

Majorando o erro em l_{20} por $\frac{1}{21}$ temos:

$$\begin{array}{rcll}
E_{19} &=& \frac{1}{20} \times \frac{1}{21} &= 0.0024 \\
E_{18} &=& \frac{1}{19} \times \frac{1}{20} \times \frac{1}{21} &= 0.00012 \\
\vdots && \vdots & \\
E_{15} &=& 4 \times 10^{-8}. &
\end{array}$$

Continuando o algoritmo, obtemos:

$$\begin{array}{lll} l_8 \simeq 0.100\,932\,0 & & l_4 \simeq 0.170\,893\,4 \\ l_7 \simeq 0.112\,383\,5 & & l_3 \simeq 0.207\,276\,7 \\ l_6 \simeq 0.126\,802\,4 & & l_2 \simeq 0.264\,241\,1 \\ l_5 \simeq 0.145\,532\,0 & & l_1 \simeq 0.367\,879\,5 \end{array}$$

Exemplo 3

Vamos agora considerar o problema de calcular a média aritmética de dois números a e b.

Algoritmo 1	Algoritmo 2	Algoritmo 3
1) Entrada(a, b)	1) Entrada(a, b)	1) Entrada(a, b)
2) $s \leftarrow a + b$	2) $s_1 \leftarrow \frac{a}{2}$	2) $d_1 \leftarrow a - b$
3) $m \leftarrow \frac{s}{2}$	3) $s_2 \leftarrow \frac{b}{2}$	3) $d_1 \leftarrow \frac{d_1}{2}$
4) Saída(m)	4) $m \leftarrow s_1 + s_2$	4) $m \leftarrow d_1 + b$
	5) Saída(m)	5) Saída(m)

Para um dado $F = F(b, n, l_1, l_2)$ podemos ter conforme os valores $a, b \in F$ os seguintes problemas:

- Algoritmo 1: *overflow* em 2
- Algoritmo 2: *underflow* em 2 e 3

No algoritmo 3 não teremos provavelmente nenhum erro operacional, mas poderemos ter um erro no comando (2) se houver cancelamento subtrativo.

Exercícios

i. Utilizando a expansão de Taylor de 15^a para a função exponencial, conforme equação (2.2), obtenha as entradas da Tabela 2.5. Utilize uma plataforma para computação numérica tal como Matlab ou Octave.

ii. Calcule l_n, $n = 1, \ldots, 15$, utilizando as expressões (2.3) e (2.4).

2.7 Instabilidade de Problemas

Na seção anterior investigamos a instabilidade inerente a algoritmos. Aqui, nos concentramos nos problemas que geram instabilidade intrínseca.

Considere a tarefa de equilibrar um lápis, conforme a Figura 2.2. A segunda tarefa (equilíbrio) é instável pois, se o lápis ficar de pé, será por algumas frações de segundo e depois cairá. Já, no caso estável, uma pequena perturbação na posição do lápis não acarretará a queda, voltando este a posição de equilíbrio. Algo semelhante ocorre com problemas numéricos.

Figura 2.2: Ilustração de equilíbrio estável e instável.

Exemplo

Consideremos então o problema de encontrar as raízes do polinômio:

$$\begin{aligned} p(x) &= x^{20} - 210x^{19} + \dots \\ &= (x-1)(x-2)(x-3)\dots(x-19)(x-20) \\ &= 0. \end{aligned} \tag{2.5}$$

Obviamente as raízes são $1, 2, 3, \dots 20$ e estão bem separadas. Computando as raízes de $p(x) + 2^{-23}x^{19} = 0$ com Matlab[1] obteremos:

$R_1 = 1.000\ 000\ 000$
$R_2 = 1.999\ 999\ 999$
$R_3 = 3.000\ 000\ 000$
$R_4 = 3.999\ 999\ 999$
$R_5 = 5.000\ 000\ 072$
$R_6 = 5.999\ 993\ 056$
$R_7 = 7.000\ 303\ 398$
$R_8 = 7.993\ 025\ 044$
$R_9 = 9.147\ 281\ 378$
$R_{10} = 9.502\ 011\ 297$
$R_{11}, R_{12} = 10.892\ 998\ 111 \pm 1.149\ 333\ 128i$
$R_{13}, R_{14} = 12.821\ 708\ 789 \pm 2.123\ 455\ 162i$
$R_{15}, R_{16} = 15.305\ 903\ 612 \pm 2.775\ 365\ 983i$
$R_{17}, R_{18} = 18.181\ 314\ 032 \pm 2.548\ 942\ 153i$
$R_{19}, R_{20} = 20.476\ 768\ 271 \pm 1.039\ 017\ 467i$

[1]Utilizamos as seguintes diretivas: *syms x*; $p = (x-1)*(x-2)*(x-3)*(x-4)*(x-5)*(x-6)*(x-7)*(x-8)*(x-9)*(x-10)*(x-11)*(x-12)*(x-13)*(x-14)*(x-15)*(x-16)*(x-17)*(x-18)*(x-19)*(x-20)+2^(-23)*x^19$; *solve*(p)

Note que um termo da equação mudou de $-210x^{19}$ para $-210x^{19} + 2^{-23}x^{19}$, ou seja, uma mudança no vigésimo dígito da base 2 de um dos coeficientes. Apesar desta pequena perturbação, o resultado é completamente inesperado e as mudanças nas raízes são grandes. A razão desta mudança drástica não é o arredondamento nem o algoritmo e sim um problema de condicionamento.

Há certos problemas que, quando sofrem alteração nos dados de entrada, têm na sua resposta uma pequena diferença proporcional, enquanto outros mostram grande variação no resultado mesmo com uma pequeníssima alteração nos dados de entrada. Os primeiros problemas são ditos bem condicionados e os segundos são ditos mal condicionados. A noção de problemas bem e mal condicionados será elaborada na parte de solução de sistemas de equações lineares.

2.8 Exercícios

Exercício 2.1 Seja P um problema, F um sistema de ponto flutuante, e A um algoritmo numérico para resolver P. Prof. Kunz afirma que a exatidão do algoritmo A aumenta se aumentarmos a precisão da máquina. Você discorda ou concorda com Prof. Kunz? Justifique sua resposta.

Exercício 2.2 Seja P um problema e F um sistema de ponto flutuante. Sejam ainda A_1 e A_2 dois algoritmos distintos para resolver o problema P. Sabemos que A_1 e A_2 resolvem P. Sem conhecer o funcionamento de A_1 e A_2, Prof. Gillet afirma que há duas possibilidades:

a) A_1 é mais exato e mais rápido do que A_2; ou

b) A_2 é mais exato e mais rápido do que A_1. Se você concorda com o Prof. Gillet, indique como que se pode descobrir qual algoritmo é melhor. Se você discorda, desenvolva uma justificativa.

Exercício 2.3 Dê a definição (sucinta) de erro inerente, de discretização e de arredondamento. Individualmente, como se pode tratar cada um destes tipos de erro?

Exercício 2.4 Sejam $F_1 = (b_1, n_1, e_{11}, e_{21})$ e $F_2 = (b_2, n_2, e_{12}, e_{22})$ dois sistemas de ponto flutuante. Sabemos que $b_1 < b_2$. Seja $x \in \mathbb{R}$ um número real, x_1 a representação de x em F_1, e x_2 a representação de x em F_2. Prof. Tellis afirma que para n_1 grande o suficiente, $|x_1 - x| \leqslant |x_2 - x|$. Você concorda ou discorda do Prof. Tellis? Justifique a sua resposta.

Exercício 2.5 Seja P um problema numérico e A um procedimento para resolver P. O procedimento A possui erros, por isso não é um algoritmo. Seja $x = A(x_0)$ a saída produzida pelo procedimento quando este recebe uma estimativa inicial x_0 da solução como entrada. Seja x^* uma solução para P. A probabilidade de $A(x_0) = x^*$ é dada por $p = 1/3$, sendo esta independente de x_0. O Prof. Beans afirma que podemos utilizar o procedimento na busca de uma solução de P? Se você concorda, indique como utilizar A. Caso contrário, justifique a impossibilidade.

Exercício 2.6 Considere o sistema de ponto flutuante $F = (2, 3, -5, 5)$. Quantas soluções admite a equação $1 + x = 1$, onde $x \in F$? Utilize arredondamento por falta.

Exercício 2.7 Assumindo precisão infinita, converta os números abaixo da base binária para a base decimal, ou da base decimal para a base binária, conforme indicação:

$$\begin{aligned} (11100.1101)_2 &= (\qquad\qquad)_{10} \\ (0.011011)_2 &= (\qquad\qquad)_{10} \\ (67)_{10} &= (\qquad\qquad)_2 \\ (93.125)_{10} &= (\qquad\qquad)_2 \end{aligned}$$

Exercício 2.8 Dado o sistema de ponto flutuante $F = F(10, 8, -99, 99)$, represente os números abaixo neste sistema de ponto flutuante:

$$\begin{aligned} x_1 &= (1043.625)_{10} \\ x_2 &= (0.0000415)_{10} \\ x_3 &= (213.013)_4 \\ x_4 &= (0.00101)_2 \end{aligned}$$

Exercício 2.9 É possível existir um sistema de ponto flutuante com $e_1 = -2$, $e_2 = 5$, e $n = 2$ com 37 elementos? Se sim, qual a base deste sistema? Caso contrário, qual é o menor número de elementos que podemos ter com este sistema?

Exercício 2.10 Considere a integral definida $y_n = \int_0^2 n^{n \log_n x} e^{x-1} dx$. Obtenha uma fórmula recursiva tal que y_n seja uma função de y_{n-1}: $y_n = f(y_{n-1})$. Calcule analiticamente y_1, depois obtenha y_2, y_3 e y_4 usando a fórmula recursiva.

Exercício 2.11 Seja $F(b, n, e_1, e_2) = F(2, 3, -1, 2)$ um sistema de ponto flutuante.

i. Encontre a região de underflow.

ii. Encontre a região de overflow.

iii. Qual é o menor elemento x de F?

iv. Para $x = 1/3$, encontre ∇x.

v. Para $x = 1/3$, encontre Δx.

Exercício 2.12 Seja $x = 0.b_1b_2 \ldots b_k$ a representação ponto flutuante de um número y. Qual a razão para que b_1 seja maior do que zero?

Exercício 2.13 Responda as questões abaixo:

i. Seja x um número real e $F(b, n, e_1, e_2)$ um sistema de ponto flutuante. Seja $\delta^{max} = \max\{x \in F\}$ e $\delta^{min} = \min\{x \in F\}$. Suponha que x está dentro da região de representação de F, ou seja, $\delta^{min} \leqslant x \leqslant \delta^{max}$. Se adotarmos ∇x como função de arredondamento, podemos afirmar que o erro absoluto $|\nabla x - x|$ é limitado?

ii. Seja x um número real e $F(b, n, e_1, e_2)$ um sistema de ponto flutuante. Seja $\delta^{max} = \max\{x \in F\}$ e $\delta^{min} = \min\{x \in F\}$. (i) Se x está dentro da região de representação de F, ou seja, $\delta^{min} \leqslant x \leqslant \delta^{max}$, (ii) se adotarmos ∇x como função de arredondamento e (iii) se $x \neq 0$, então podemos afirmar que o erro relativo definido por $|\nabla x - x|/x$ é limitado?

iii. É sempre possível implementar ox como uma função de ∇x e Δx?

iv. Assuma que $x \in \mathbb{R}$ não está na região de overflow de um sistema de ponto flutuante $F(b, n, e_1, e_2)$. Considere o conjunto $S = \{x \in \mathbb{R} : x \notin$ região de overflow, $\nabla x = \Delta x\}$. Podemos afirmar que S é finito?

Exercício 2.14 Para a função $f(x) = \sqrt{x}\sin(x)$ execute as tarefas abaixo:

a) Obtenha uma aproximação $\tilde{f}(x)$ em torno do ponto $x = \pi/2$ com uma série de Taylor de segunda ordem. (O erro deve ser da ordem $O(|\Delta x|^3)$).

b) Calcule $\tilde{f}(\pi/3)$.

c) Calcule o erro relativo cometido ao adotarmos o valor $\tilde{f}(\pi/3)$ em vez de $f(\pi/3)$.

Exercício 2.15 Converta o número $(0.132)_4$ na base 4 para o seu correspondent na base 5 utilizando até 5 dígitos na mantissa.

Capítulo 3

Resolução de Equações Não-Lineares

3.1 Introdução

Dada uma função $f : \mathbb{R} \to \mathbb{R}$, um problema de grande interesse é a determinação da existência e cálculo de uma raiz x de f, ou seja, x tal que $f(x) = 0$. O primeiros estudos datam do Século IX quando os trabalhos de matemáticos árabes que difundiram a utilização do sistema decimal e do zero na escrita de números. Com Bhaskara, tornou-se conhecida a fórmula para resolução de equações quadráticas: $f(x) = ax^2 + bx + c$. Mais tarde, os matemáticos Nicolo Fontana e Jerônimo Cardano desenvolveram métodos para solução de equações de 3^o e 4^o grau. O matemático Niels Abel provou que as equações de quinto grau ou superior não podem ser resolvidas de uma forma coerente.

Outro resultado relevante é o Teorema Fundamental da Álgebra, enunciado por D'Alembert em 1746: "toda a equação polinomial de grau n possui exatamente n raízes" foi demonstrado por Gauss em 1799. A partir daí, até os dias atuais, os métodos de cálculo das n raízes de um polinômio de grau n são voltados aos métodos iterativos. Tais métodos são também aplicáveis às equações transcedentes[1]. Os principais métodos iterativos para cálculo das raízes de equações algébricas ou transcedentes são de três tipos:

Métodos de quebra: Para aplicarmos os métodos de quebra temos que ter um intervalo $[a, b]$ onde a função troca de sinal. Partimos o intervalo em dois outros intervalos e verificamos qual contém a raiz desejada e assim prosseguimos.

[1]São funções que não podem ser definidas diretamente através de fórmulas algébricas como, por exemplo, funções exponenciais, logarítmicas e trigonométricas.

Métodos de ponto fixo: Começamos de uma aproximação inicial x_0 e construímos uma seqüência $\{x_j\}_{j=1}^{n}$ na qual cada termo é dado por $x_{j+1} = g(x_j)$, onde g é uma função de iteração. Conforme as propriedades de g, surgem diferentes tipos de métodos de ponto fixo.

Métodos de múltiplos passos: Estes métodos constituem uma generalização do anterior onde para determinar um ponto x_{j+1} utilizamos vários pontos anteriores: $x_j, x_{j-1}, \ldots, x_{j-p}$. Com a abordagem iterativa se faz necessário determinar um intervalo inicial, um ou mais pontos para construirmos a seqüência $\{x_j\}$ e, mediante certas condições, teremos que a raiz x^* será dada por

$$x^* = \lim_{j \to \infty} x_j$$

3.1.1 Exemplo de Aplicação

Aqui ilustramos a aplicação do problema de encontrar a raiz de uma equação não-linear em circuitos elétricos. O circuito elétrico exemplo, descrito na Figura 3.1, consiste de uma fonte de tensão contínua E em série com um resistor e um elemento não-linear, cuja queda de tensão x é dada por uma função não linear $g(x)$ que regula a corrente y que circula na malha, sendo y uma função da queda de tensão no componente não linear, i.e., $y = g(x)$. Aplicando a lei das malhas no circuito obtemos

$$\begin{aligned} E &= Ry + x & (3.1)\\ &= Rg(x) + x & (3.2) \end{aligned}$$

portanto

$$g(x) - \frac{E}{R} + \frac{x}{R} = 0. \quad (3.3)$$

O problema é encontrar um ponto de operação para o sistema, ou seja, valores de tensão ou corrente que definam um ponto de equilíbrio que nada mais é do que a solução de (3.3). Fazendo

$$f(x) = g(x) - \frac{E - x}{R}$$

o problema é encontrar um zero de $f(x)$. Para o caso do gráfico dado na Figura 3.1, observamos que a função $f(x)$ tem três possíveis soluções, cada uma delas definindo um ponto de operação distinto para o circuito.

O número e existência de solução de uma função não-linear $f(x)$ depende intrinsecamente da função. Alguns exemplos são:

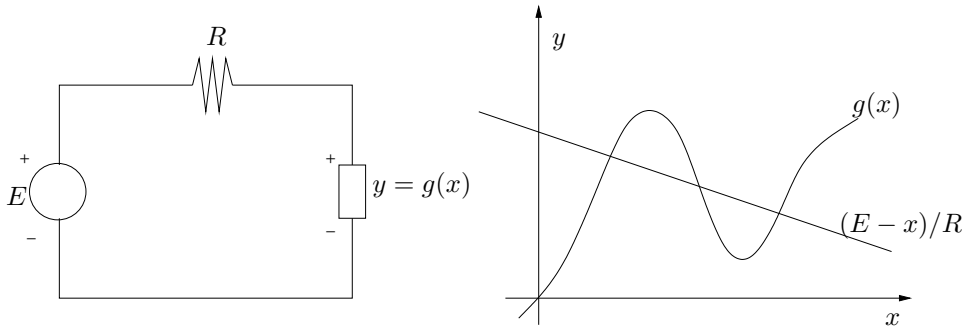


Figura 3.1: Circuito elétrico com componente não-linear.

a) $f(x) = e^x$ não tem zero;

b) $f(x) = e^x - e^{-x}$ tem apenas um zero;

c) $f(x) = e^x - e^{-x} - 3x$ tem três zeros; e

d) $f(x) = \cos x - 1/2$ tem um número infinito de zeros.

Ilustrações gráficas das funções acima estão na Figura 3.2.

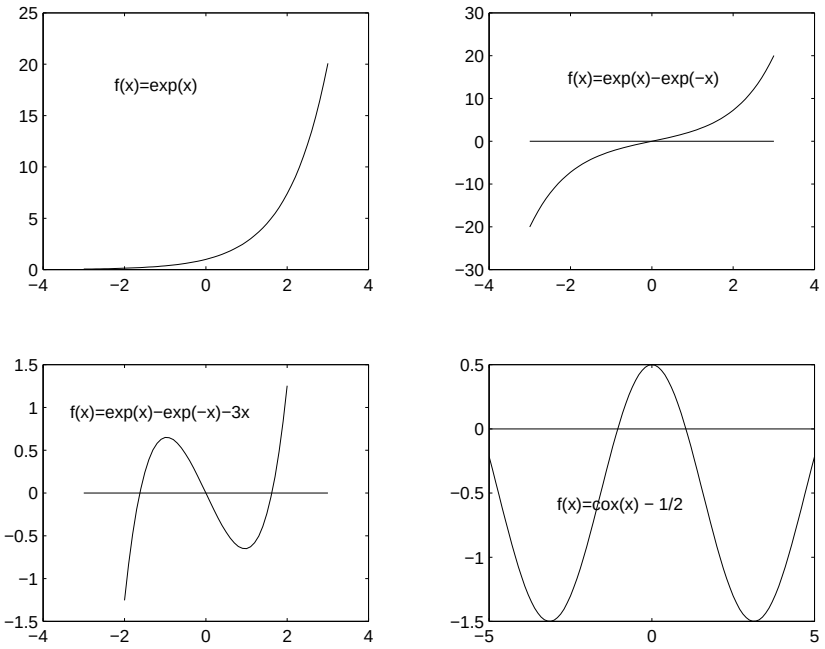


Figura 3.2: Número de zeros para várias funções.

3.2 Ordem de Convergência

O conceito de convergência é muito importante no estudo da matemática numérica e será utilizado nas próximas seções. Em geral este assume expressões como:

a) o método converge como $1/n$;

b) o método converge como $1/k^{3.5}$;

c) o método converge como h^3;

d) o erro é da ordem h^4;

e) o erro decai como e^{-n}; e

f) a taxa de convergência é $\log(n)/n$.

Precisamos de uma noção sobre como o erro tende a zero à medida que os parâmetros variam. Isto nos permite comparar diferentes métodos.

Definição 3.1 *Uma seqüência $x_0, x_1, \ldots$ converge para $\bar{x}$ se dado $\epsilon > 0$ existe um I tal que qualquer que seja $i > I$, $|x_i - \bar{x}| < \epsilon$. Neste caso, temos:*

$$\lim_{j \to \infty} x_j = \bar{x}$$

Definição 3.2 *Seja uma seqüência $x_0, x_1, \ldots$ que converge para $\bar{x}$. Seja $e_j = |x_j - \bar{x}|$. Se existe um número $p \geqslant 1$ e uma constante $c \neq 0$ tal que*

$$\lim_{j \to \infty} \frac{e_{j+1}}{e_j^p} = c$$

então p é dito ordem de convergência *e c, constante assintótica de erro.*

Se $p = 1, 2$, ou 3, então a convergência é dita linear, quadrática ou cúbica. Se a seqüência $x_0, x_1, \ldots$ é produzida por uma função ϕ onde

$$x_{j+1} = \phi(x_j, x_{j-1}, \ldots, x_{j-m+1})$$

então dizemos que ϕ é de ordem m. Quando a convergência é linear, então isto significa que a cada passo do método o erro é reduzido (aproximadamente) de um fator constante. Se a convergência é quadrática, então o erro é assintoticamente reduzido do quadrado do erro anterior.

Exercícios

Considere as séries:

i. $x_k = 1/k$, $k = 1, 2, \ldots$;

ii. $x_k = 1/2^k$, $k = 1, 2, \ldots$; e

iii. $x_{k+1} = \alpha^k x_k + \beta$, $k = 1, 2, \ldots$, com $\alpha \in (0, 1)$, $\beta \in \mathbb{R}$ e $x_1 \in \mathbb{R}$.

Para cada série, você pode afirmar se ela converge? Utilize a definição de convergência para responder a pergunta. Se for convergente, qual é a taxa de convergência?

3.3 Métodos Iterativos para Resolução de Equações

Qualquer método iterativo é formado de quatro partes:

a) Estimativa inicial: uma ou mais aproximações para a raiz desejada.

b) Atualização: uma fórmula que atualize a solução aproximada.

c) Critério de parada: uma forma de estabelecer quando parar o processo iterativo em qualquer caso.

d) Estimador de exatidão: está associado ao critério de parada e provê uma estimativa do erro cometido.

O item (a) é obtido via separação das raízes, enquanto o item (b) será analisado posteriormente caso a caso conforme o método. Vejamos os itens (c) e (d) simultaneamente. Em princípio, podemos parar um processo iterativo de quatro maneiras:

a) $\frac{|x_j - x_{j-1}|}{|x_j|} < \epsilon_1$

b) $|f(x_j)| < \epsilon_2$

c) $DIGSE(x_j, x_{j-1}) \geqslant k$

d) $i > L$

onde: ϵ_1, ϵ_2 são valores de tolerância dados, k é o número de dígitos significativos exatos requeridos na aproximação final, e L é o número máximo de iterações permitido.

O critério (a) se refere ao erro relativo entre x_j e x_{j-1}. O critério (b) é um limite para o módulo do valor de f em x_j. Pode ocorrer que um critério seja satisfeito sem que os outros sejam, conforme é ilustrado na Figura 3.3. Para certos valores de ϵ_1 e ϵ_2, o segundo critério pode ser satisfeito na Figura 3.3(a) enquanto o primeiro critério pode falhar. Na parte (b) da figura o inverso ocorre.

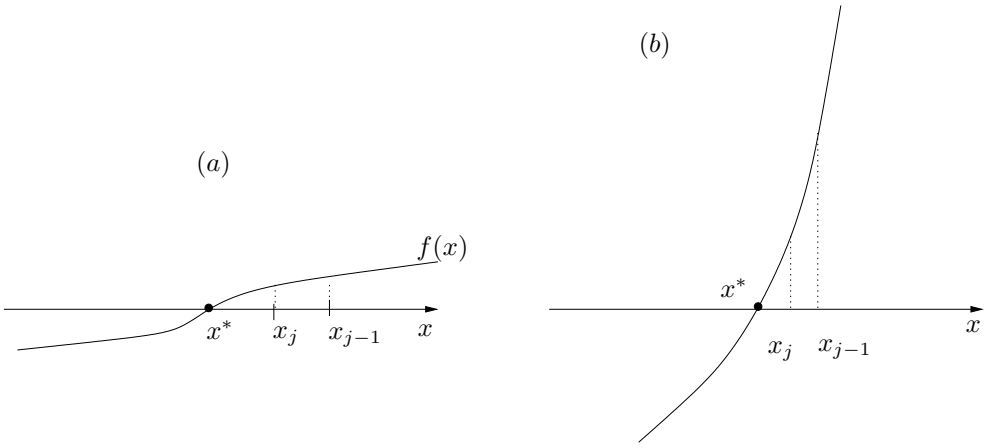


Figura 3.3: Dificuldades com critérios de convergência.

O terceiro critério está relacionado com o erro relativo, porém dá uma idéia mais clara da exatidão obtida.

3.4 Métodos de Quebra

Os métodos de quebra são os mais intuitivos geometricamente, contudo, são os que convergem mais lentamente. A partir de um intervalo que contenha uma raiz para $f(x) = 0$ se parte este intervalo em outros menores que ainda contenham pelo menos uma raiz de $f(x) = 0$. É necessário que f troque de sinal no intervalo inicial e seja contínua dentro do intervalo.

3.4.1 Método da Bisecção

O método da bisecção, inspirado no Teorema de Bolzano, parte de um intervalo $[a, b]$ que contenha uma raiz para $f(x) = 0$ onde $f(a) \cdot f(b) < 0$,

ou seja, $f(x)$ corta o eixo x em pelo menos um ponto de $[a, b]$. O passos do algoritmo são dados a seguir.

Passos Gerais do Algoritmo

1) Calcula-se $f(x)$ no ponto médio de $[a, b]$: $x_m = \frac{a+b}{2}$

2) Se $f(x_m) \neq 0$ (i.e., $f(a) \cdot f(x_m) < 0$ ou $f(x_m) \cdot f(b) < 0$), escolhe-se um novo intervalo de modo que f tenha sinais opostos nas extremidades

3) Repete-se a partir de (1) até que tenhamos chegado "suficientemente perto da raiz"

Sejam $x_0 = a$ e $x_1 = b$ dois pontos que definem um intervalo l_0. Da Figura 3.4 observa-se que $f(a) \cdot f(x_m) < 0$ onde $x_m = \frac{a+b}{2} = x_2$. Portanto, determinamos que o intervalo $l_1 = [x_0, x_2]$ contém uma raiz de $f(x) = 0$. Mais detalhadamente, o método da bisecção pode ser descrito conforme o algoritmo abaixo. A Figura 3.4 ilustra o funcionamento do algoritmo.

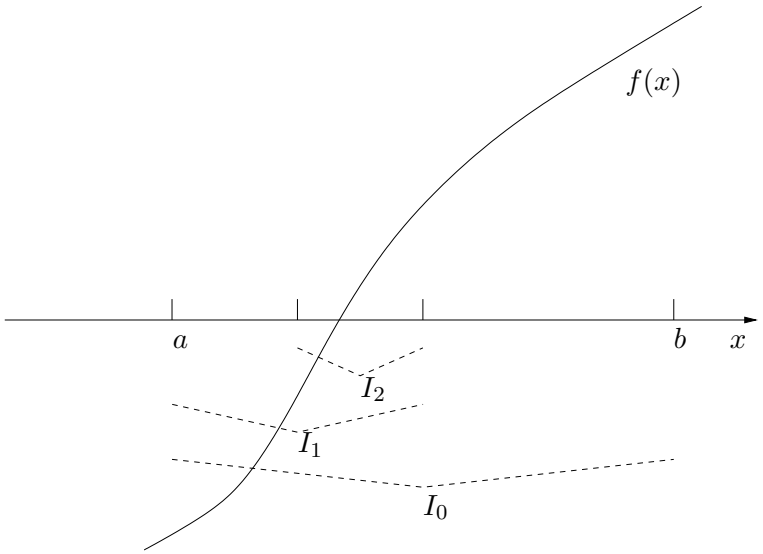


Figura 3.4: Ilustração do comportamento do algoritmo da bisecção

Bisecção$(f, a, b, L, \epsilon_1, \epsilon_2)$
1: $x_0 \leftarrow a$, $x_1 \leftarrow b$, $i \leftarrow 0$
2: $f_0 \leftarrow f(x_0)$, $f_1 \leftarrow f(x_1)$
3: **if** $f_0 \cdot f_1 > 0$ **then**
4: Saída "erro: não $[a, b]$ não satisfaz condições iniciais"

end if
while $|f_0| > \epsilon_2$ e $|f_1| > \epsilon_2$ e $i \leq L$ **do**
  if $|x_0 - x_1| < \epsilon_1 |x_1|$ **then**
    Saída (x_0, x_1);
  end if
  $x_2 \leftarrow (x_0 + x_1)/2$
  $f_2 \leftarrow f(x_2)$
  if $f_2 \cdot f_0 < 0$ **then**
    $x_1 \leftarrow x_2$
    $f_1 \leftarrow f_2$
  else
    $x_0 \leftarrow x_2$
    $f_0 \leftarrow f_2$
  end if
  $i \leftarrow i + 1$
end while
if $i > L$ **then**
  Saída "não atingiu exatidão em L iterações"
end if
if $|f_0| \leqslant \epsilon_2$ **then**
  Saída (x_0)
else
  Saída (x_1)
end if

A busca de um intervalo inicial $[a, b]$ onde $f(a).f(b) < 0$ pode ser exemplificada para a função $f(x) = e^{-x} - x$ que está ilustrada na Figura 3.5. A partir dos valores de $f(x)$ obtidos na Tabela 3.1, podemos deduzir que o intervalo inicial $[a, b] = [0.5, 1.0]$ satisfaz a condição necessária para aplicação do método da bisecção. Conforme dados da tabela, podemos fazer $a = 0.5$ e $b = 1.0$ tal que $[a, b]$ contém uma raiz para $f(x)$.

Tabela 3.1: Valores da função $f(x) = e^{-x} - x$

x	$f(x)$
0	1
0.5	0.1065
1.0	-0.6321
1.5	-1.2768
2.0	-1.8646

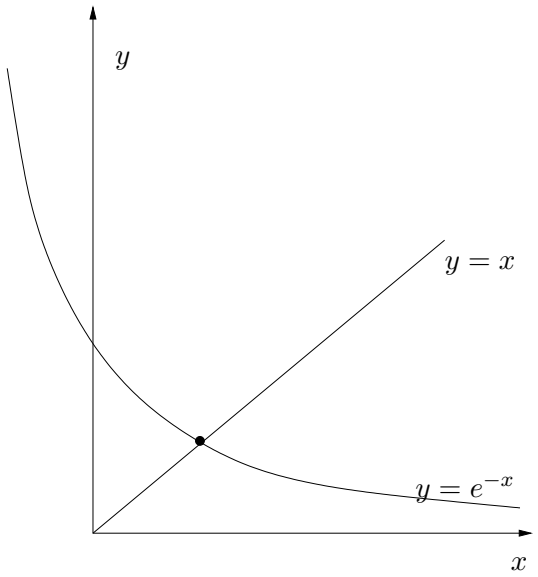


Figura 3.5: Função $f(x) = e^{-x} - x$.

Observações:

1) pode ser difícil encontrar um intervalo inicial, conforme ilustração da Figura 3.6; e

2) se ocorrer um erro de arredondamento, mesmo que pequeno, no momento que se avalia o sinal do ponto médio, o intervalo resultante pode não conter uma raiz.

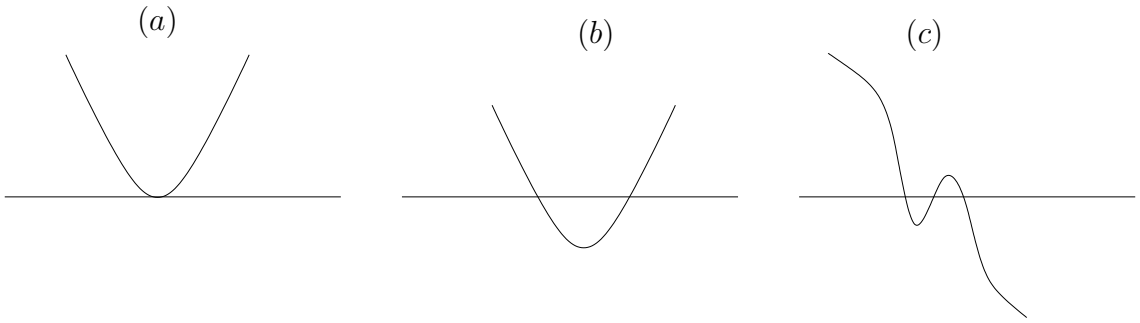


Figura 3.6: Dificuldade na obtenção do intervalo inicial

Ordem de Convergência da Bisecção

Seja $\bar{x}$ a raiz de $f(x)$. Seja $e_j = |x_j - \bar{x}|$ o erro da iteração j. Uma vez que $x_{j+1} = \frac{x_j + x_{j-1}}{2}$ tal que $f(x_j)f(x_{j-1}) < 0$, temos que

$$e_{j+1} \leqslant \frac{e_j}{2} \Leftrightarrow \frac{e_{j+1}}{e_j} \leqslant \frac{1}{2} \tag{3.4}$$

Logo,

$$\lim_{j\to\infty} \frac{e_{j+1}}{e_j} = \frac{1}{2}$$

Portanto, fica evidenciado que o método da bisecção tem convergência linear.

Vamos agora relacionar a convergência linear com o número $DIGSE$, obtido a cada iteração:

$$\begin{aligned} &DIGSE(x_{j+1}, x_j) - DIGSE(x_j, x_{j-1}) = \\ &\quad = -\left(0.3 + \log\left(\tfrac{|x_{j+1}-x_j|}{|x_j|}\right)\right) + \left(0.3 + \log\left(\tfrac{|x_j-x_{j-1}|}{|x_{j-1}|}\right)\right) = \\ &\quad = \log\left(\tfrac{|x_j-x_{j-1}|}{|x_{j-1}|}\right) - \log\left(\tfrac{|x_{j+1}-x_j|}{|x_j|}\right) \end{aligned}$$

Indicando $E_R(x_j) = \frac{|x_j-x_{j-1}|}{|x_{j-1}|}$ e usando (3.4), obtemos

$$\begin{aligned} DIGSE(x_{j+1}, x_j) - DIGSE(x_j, x_{j-1}) &= \log \frac{E_R(x_j)}{E_R(x_{j+1})} \\ &\geqslant \log 2 \\ DIGSE(x_{j+1}, x_j) &\approx DIGSE(x_j, x_{j-1}) + 0.33 \end{aligned}$$

A velocidade de convergência da bisecção é $0.3DIGSE/$ iteração, isto é, a cada três ou quatro iterações ganha-se um $DIGSE$, ou ainda um bit a cada iteração.

Exemplo: Minimização de Funções

Suponha que cada cm^2 de latão custa R\$ $2,00$. Qual deve ser a altura h (em cm) e o raio r (em cm) de uma lata sem tampa, de menor custo possível, tal que seu volume seja maior ou igual a $500cm^3$? Podemos expressar este problema em uma linguagem matemática conhecida como *programação matemática*:

$$\begin{aligned} P: \quad & Minimize \quad f(r,h) = 2(\pi r^2 + 2\pi rh) \\ & \underset{r,h}{} \\ & Sujeito\ a: \\ & \qquad \pi r^2 h \geqslant 500 \end{aligned}$$

Observando que $h = 500/(\pi r^2)$ e substituindo esta expressão em P, obtemos:

$$P: \quad \underset{r,h}{Minimize} \quad f(r,h) = 2(\pi r^2 + 2\pi rh) = 2\left[\pi r^2 + 2\pi r \tfrac{500}{\pi r^2}\right] = 2\pi r^2 + \tfrac{4\times 500}{r}$$

Note que f é uma função convexa e, portanto, o ponto de mínimo pode ser encontrado fazendo $\frac{df}{dr} = 0$, o que nos leva a:

$$\frac{df}{dr} = 4\pi r - \frac{2000}{r^2} = 0$$

portanto, o raio ótimo é:

$$r^* = \left(\frac{500}{\pi}\right)^{\frac{1}{3}}$$

e a altura ótima é:

$$h^* = \frac{500}{\pi\left(\frac{500}{\pi}\right)^{\frac{2}{3}}} = \frac{500^{\frac{1}{3}}}{\pi^{\frac{1}{3}}} = \left(\frac{500}{\pi}\right)^{\frac{1}{3}} = r^*$$

Em valores numéricos: $(r^*, h^*) = (5.4193, 5.4193)$.

Tarefa: utilize o método da bisecção para encontrar a solução ótima (r^*, h^*).

Exemplo: Autovalores

Encontre um autovalor para a matriz A dada por:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4 & -1 & 5 & 1 \end{bmatrix}$$

Os autovalores de uma matriz $A \in \mathbb{R}^{n\times n}$ são definidos como as raízes da equação característica de grau n:

$$p(\lambda) = det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 = 0$$

Assim, podemos obter a equação característica de A como segue:

$$p(\lambda) = det \begin{bmatrix} \lambda & -1 & 0 & 0 \\ 0 & \lambda & -1 & 0 \\ 0 & 0 & \lambda & -1 \\ 4 & 1 & -5 & \lambda - 1 \end{bmatrix}$$

Fazendo uso da fórmula recursiva para cálculo de determinante, deduzimos que:

$$\begin{aligned} p(\lambda) &= \lambda(-1)^{1+1} det \begin{bmatrix} \lambda & -1 & 0 \\ 0 & \lambda & -1 \\ 1 & -5 & \lambda - 1 \end{bmatrix} + 4(-1)^{4+1} det \begin{bmatrix} -1 & 0 & 0 \\ \lambda & -1 & 0 \\ 0 & \lambda & -1 \end{bmatrix} \\ &= \lambda[\lambda^2(\lambda - 1) + 1 - 5\lambda] - 4[-1] \\ &= \lambda^4 - \lambda^3 + \lambda - 5\lambda^2 + 4 \end{aligned}$$

Fazendo $p(\lambda) = 0$ e obtendo as raízes, podemos verificar que $\sigma(A) = \{2.5616, -1.5616, -1, 1\}$ é o conjunto de autovalores de A. As raízes de $p(\lambda)$ são ilustradas na Figura 3.7.

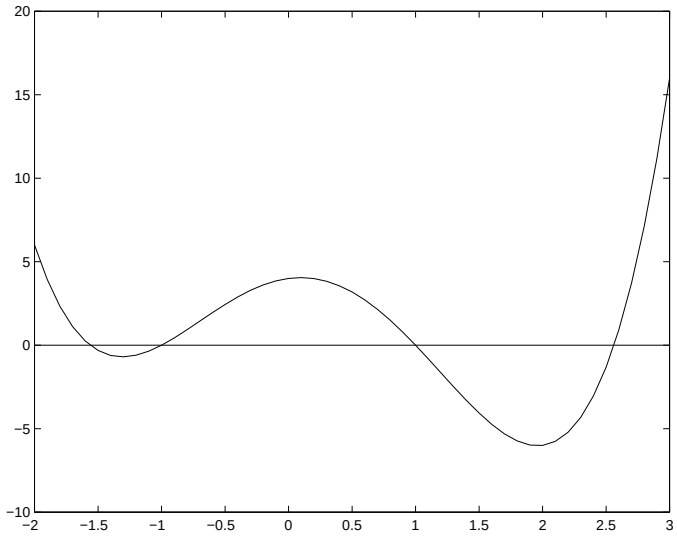


Figura 3.7: Raízes do polinômio característico $p(\lambda) = \lambda^4 - \lambda^3 + \lambda - 5\lambda^2 + 4$ de A.

3.4.2 Método da Falsa Posição

Sob as mesmas condições iniciais do método da bisecção, temos ainda várias maneiras de particionar o intervalo $[a, b]$ que contém pelo menos uma raiz. Por exemplo, podemos particionar $[a, b]$ na intersecção da reta que une os pontos $(a, f(a))$ e $(b, f(b))$ com o eixo x. Seja x_S tal ponto. Escolhe-se então o novo subintervalo conforme a variação do sinal da curva f. Este procedimento está ilustrado na Figura 3.8. Da intersecção da reta que une

os pontos $(a, f(a))$ e $(b, f(b))$ com a reta $y = 0$, temos que:

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

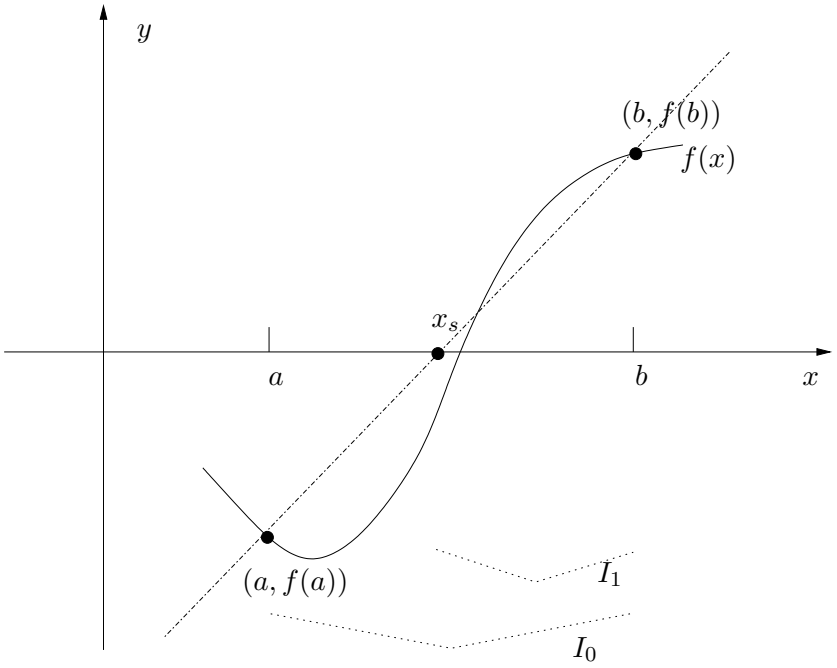


Figura 3.8: Ilustração do método da falsa posição

Falsa_Posição(a, b, f, k, Lim)

1: $x_1 \leftarrow a$ e $x_2 \leftarrow b$
2: $f_1 \leftarrow f(x_1)$ e $f_2 \leftarrow f(x_2)$
3: $x_a \leftarrow a$, $x_S \leftarrow b$, e $i \leftarrow 1$
4: **while** $i \leq Lim$ e $DIGSE(x_S, x_a) < k$ **do**
5: $\quad x_a \leftarrow x_S \qquad \triangleright\ x_a$ é x_S da iteração anterior
6: $\quad x_S = x_1 - \frac{(x_2 - x_1)f_1}{f_2 - f_1}$
7: $\quad f_s = f(x_S)$
8: $\quad$ **if** $f_S.f_1 < 0$ **then**
9: $\quad\quad x_2 \leftarrow x_S$
10: $\quad\quad f_2 \leftarrow f_S$
11: $\quad$ **else**
12: $\quad\quad x_1 \leftarrow x_S$
13: $\quad\quad f_1 \leftarrow f_S$
14: $\quad$ **end if**

15: $i \leftarrow i + 1$
16: **end while**
17: Saída $\{x_S, DIGSE(x_S, x_a)\}$

Exemplo 1

A tarefa consiste em encontrar um zero para o polinômio $p(x) = x^4 - 2x^3 - 6x^2 + 2$ onde $a = -1$ e $b = 0$. Note que $p(a) = -1$ e $p(b) = 2$, portanto o intervalo inicial satisfaz a condição necessária para aplicação do algoritmo da falsa posição. A Tabela 3.2 abaixo ilustra o comportamento do algoritmo para a seqüência dos 5 primeiros iterandos.

Tabela 3.2: Aplicação do algoritmo da falsa posição

i	a	x_S	$f(x_S)$
0	-1.0	$+0.0$	$+0.123$
1	-1.0	$-0.666\ 666\ 666\ 7$	-2.74×10^{-2}
2	$-0.703\ 296\ 703\ 3$	$-0.666\ 666\ 666\ 7$	-1.95×10^{-4}
3	$-0.696\ 650\ 702\ 1$	$-0.666\ 666\ 666\ 7$	-1.34×10^{-6}
4	$-0.609\ 603\ 305\ 1$	$-0.666\ 666\ 666\ 7$	-9.22×10^{-9}
5	$-0.696\ 602\ 978\ 6$	$-0.666\ 666\ 666\ 7$	-2.22×10^{-10}

Exemplo 2

Aqui vamos aplicar o método da falsa posição para encontrar uma raiz de $f(x) = x^4 - 14x^2 + 24x - 10$. Conforme Figura 3.9 existe uma raiz no intervalo $I = [-5, 0]$. As Figuras 3.10 e 3.11 ilustram o comportamento do método da falsa posição nas 8 primeiras iterações. O método converge para a raiz $x^* = -4.4593$.

Exercícios

i. Considere um método de quebra que, ao contrário do método da bisecção que testa o ponto médio $(x_0 + x_1)/2$, testa o ponto $x_s = (x_0 + 3x_1)/4$ para fazer a secção do intervalo $[x_0, x_1]$ em $[x_0, x_s]$ (se $f(x_0)f(x_s) < 0$) ou $[x_s, x_1]$ (se $f(x_s)f(x_1) < 0$). Implemente este algoritmo em matlab. Compare o desempenho do método da bisecção e o método proposto aqui na busca das raízes do polinômio característico $p(\lambda) = \lambda^4 - \lambda^3 + \lambda - 5\lambda^2 + 4$.

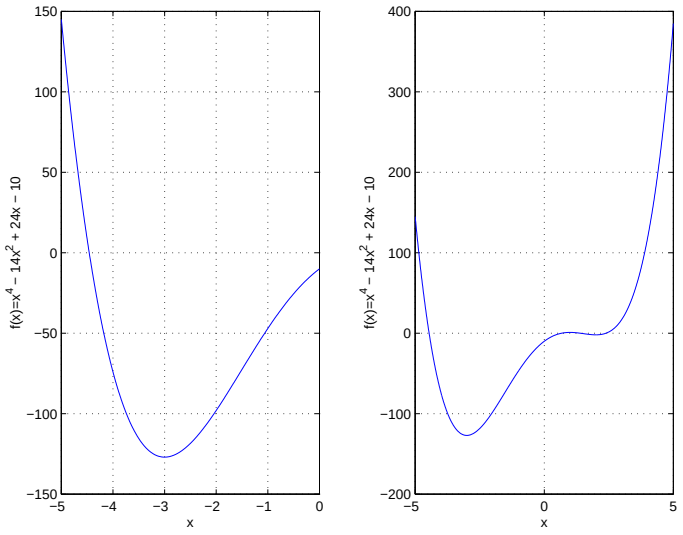


Figura 3.9: Gráfico da função $f(x) = x^4 - 14x^2 + 24x - 10$

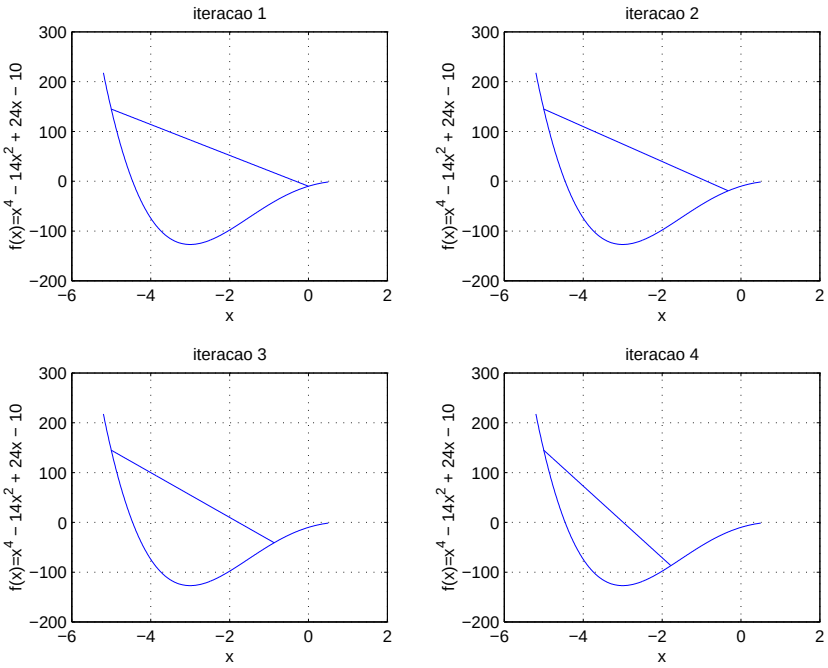


Figura 3.10: Iterações 1, 2, 3 e 4 do método da falsa posição

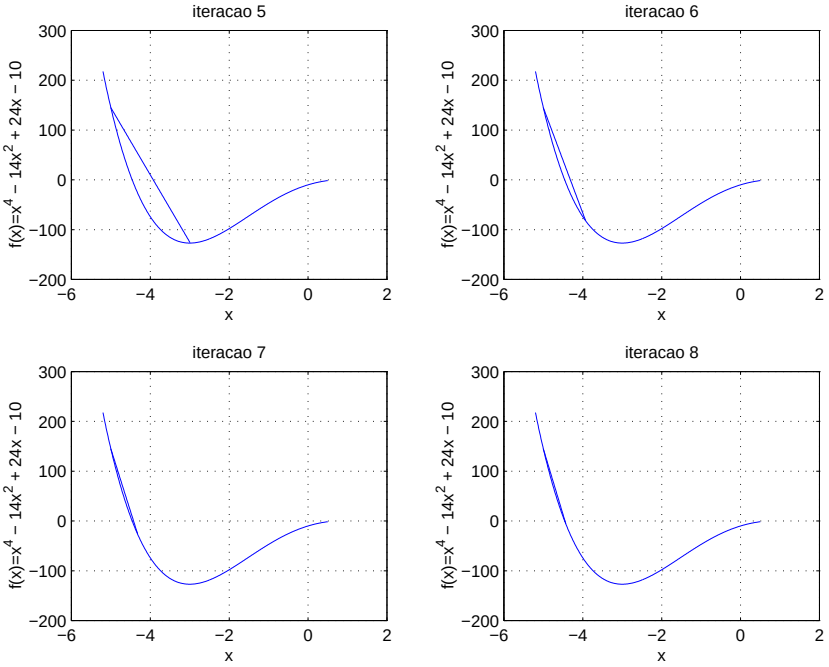


Figura 3.11: Iterações 5, 6, 7 e 8 do método da falsa posição

ii. Para o método de quebra dado no item acima, qual é a ordem de convergência do algoritmo?

3.5 Métodos de Ponto Fixo

Inicialmente relembramos que o problema de interesse é encontrar um ponto x tal que $f(x) = 0$, onde $f(x)$ é uma função não linear. Façamos a substituição

$$g(x) = x + c(x).f(x)$$

onde $c(x)$ deve ser escolhida de forma que $c(x) \neq 0$ para todo $x \in [a, b]$. Sabemos que existe uma raiz no intervalo $[a, b]$. Assim, o problema de encontrar x tal que $f(x) = 0$ pode ser transformado no problema de encontrar x tal que $g(x) = x$. Usando a função auxiliar, buscamos um *ponto fixo* x^* tal que

$$x^* = g(x^*).$$

Partindo de um ponto inicial x_0 utilizamos o processo iterativo

$$x_{k+1} = g(x_k)$$

até que convergência seja atingida para o ponto fixo x^*. Note que $x^* = g(x^*) \Leftrightarrow f(x^*) = 0$. A seguir enunciamos alguns resultados fundamentais para aplicação de métodos de ponto fixo.

Teorema 3.1 *Seja $I = [a, b]$ um intervalo e $g(x)$ uma função satisfazendo:*

a) g é contínua em I; e

b) $g(I) \subseteq I$.

Então existe pelo menos um $x^* \in I$ tal que $g(x^*) = x^*$, ou seja, g contém um ponto fixo no intervalo I.

Prova: Como $g(I) \subseteq I$, então vale

$$\begin{array}{ccccc} a & \leqslant & g(a) & \leqslant & b \\ a & \leqslant & g(b) & \leqslant & b \end{array}$$

Se $g(a) = a$ ou $g(b) = b$, então temos um ponto fixo. Se a e b não são pontos fixos, então $g(a) - a > 0$ e $g(b) - b < 0$. Considere a função

$$F(x) = g(x) - x.$$

Note que $F(x)$ é contínua com $F(a) > 0$ e $F(b) < 0$. Logo, pelo Teorema do Valor Intermediário (Teorema A.1) a função $F(x)$ deve assumir valor 0 para algum $x^* \in (a, b)$. Portanto x^* é um ponto fixo de $g(x)$ contido em I. ■

Teorema 3.2 *Seja g uma função definida em $I = [a, b]$, satisfazendo*

a) $g(I) \subseteq I$; e

b) $\forall x \in I$, $|g'(x)| \leqslant L < 1$ (isto é, g é um operador contrativo). Então existe exatamente um $x^* \in I$ tal que $x^* = g(x^*)$.

Prova: A condição (b) implica na continuidade da função g. Pelo Teorema 3.1, existe pelo menos um $x^* \in I$ tal que $x^* = g(x^*)$. Sejam x_1^* e x_2^* dois elementos distintos de I, satisfazendo $x_1^* = g(x_1^*)$ e $x_2^* = g(x_2^*)$. Daí, verificamos que:

$$\begin{array}{rcll} |x_1^* - x_2^*| & = & |g(x_1^*) - g(x_2^*)| & \\ & = & |g'(\beta)(x_1^* - x_2^*)| \text{ para algum } \beta \in [x_1^*, x_2^*] & \\ & & \text{pelo Teorema do Valor Médio} & (3.5) \\ & \leqslant & L|x_1^* - x_2^*| & \\ & < & |x_1^* - x_2^*| & \end{array}$$

Mas isto é uma contradição, portanto $x_1^* = x_2^*$. ■

Teorema 3.3 *Seja g uma função que satisfaz as condições (a) e (b) do Teorema 3.2. Para $x_0 \in I$, a seqüência $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, converge para o ponto x^* e o erro de truncamento no processo-limite cometido na k-ésima iteração satisfaz:*

$$\begin{array}{lllll} |x^* - x_k| & = & |e_{tr}| & < & \frac{L^k}{1-L}|x_1 - x_0| \qquad \textit{(erro a priori)} \\ |x^* - x_k| & = & |e_{tr}| & < & \frac{L}{1-L}|x_k - x_{k-1}| \qquad \textit{(erro a posteriori)} \end{array}$$

Prova: A partir do Teorema 3.2, sabemos que existe um ponto fixo x^* único. Sabemos também que

$$\begin{array}{rcl} |x_k - x^*| & = & |g(x_{k-1}) - g(x^*)| \\ & = & |g'(\beta)(x_{k-1} - x^*)| \text{ para } \beta \in [x_{k-1}, x^*] \\ & = & |g'(\beta)|.|x_{k-1} - x^*| \\ & \leqslant & L|x_{k-1} - x^*|. \end{array}$$

Sucessivamente, podemos verificar que

$$|x_k - x^*| \leqslant L|x_{k-1} - x^*| \leqslant L^2|x_{k-2} - x^*| \leqslant L^k|x_0 - x^*|.$$

Como $L < 1$, temos

$$\lim_{k\to\infty} L^k = 0 \Rightarrow \lim_{k\to\infty} |x_k - x^*| = 0 \Rightarrow \lim_{k\to\infty} x_k = x^*.$$

Note que

$$\begin{array}{rcl} |x_{k-1} - x_k| & = & |g(x_{k-2}) - g(x_{k-1})| = |g'(\beta)|.|x_{k-2} - x_{k-1}| \\ & \leqslant & L|x_{k-2} - x_{k-1}| \\ & \leqslant & L^2|x_{k-3} - x_{k-2}| \\ & \leqslant & \vdots \\ & \leqslant & L^{k-1}|x_0 - x_1|. \end{array}$$

Considere uma iteração k e $j > k$, podemos verificar que

$$\begin{array}{rcl} |x_j - x_k| & = & |x_j - x_{j-1} + x_{j-1} - x_{j-2} + x_{j-2} - \ldots + x_{k+1} - x_k| \\ & = & |(x_j - x_{j-1}) + (x_{j-1} - x_{j-2}) + \ldots + (x_{k+1} - x_k)| \\ & \leqslant & |x_j - x_{j-1}| + |x_{j-1} - x_{j-2}| + \ldots + |x_{k+1} - x_k| \\ & \leqslant & L^{j-1}|x_0 - x_1| + L^{j-2}|x_0 - x_1| + \ldots + L^k|x_0 - x_1| \\ & \leqslant & L^k(L^{j-1-k} + L^{j-2-k} + \ldots + L^0)|x_0 - x_1| \\ & \leqslant & L^k\left(\sum_{i=0}^{\infty} L^i\right)|x_0 - x_1| \\ & \leqslant & \frac{L^k}{1-L}|x_1 - x_0|. \end{array}$$

Fazendo $x_j \to x^*$, concluímos que

$$|e_{tr}| = |x^* - x_k| \leqslant \frac{L^k}{1-L}|x_1 - x_0|.$$

De forma semelhante, pode-se deduzir o erro de truncamento *a posteriori* conforme enunciado no teorema. ■

Teorema 3.4 *Seja g uma função definida em $I = [a, b]$, satisfazendo*

a) g é contínua;

b) $g(I) \subseteq I$; e

c) $|g(x) - g(y)| \leqslant \gamma|x - y|$ para todo $x, y \in I, x \neq y$, e algum $\gamma < 1$.

Então existe exatamente um $x^* \in I$ tal que $x^* = g(x^*)$ e o processo iterativo $x_{k+1} = g(x_k), k = 0, 1, \ldots,$ converge para x^* se $x_0 \in I$.

Prova: Primeiro, vamos demonstrar a unicidade de x^*. As condições (a) e (b) são as mesmas do Teorema 3.1, implicando na existência de pelo menos um ponto fixo em I. Suponha que existam dois pontos fixos $x_1^*, x_2^* \in I$ distintos. Então:

$$\begin{aligned} |x_1^* - x_2^*| &= |g(x_1^*) - g(x_2^*)| \\ &\leqslant \gamma|x_1^* - x_2^*| \\ &< |x_1^* - x_2^*| \end{aligned}$$

contradizendo a hipótese. Concluímos que existe exatamente um ponto fixo x^* em I.

Agora, podemos demonstrar a convergência para o ponto fixo x^*. Note que:

$$\begin{aligned} |x_k - x^*| &= |g(x_{k-1}) - x^*| \\ &= |g(x_{k-1}) - g(x^*)| \\ &\leqslant \gamma|x_{k-1} - x^*| \\ &\leqslant \quad \vdots \\ &\leqslant \gamma^k|x_0 - x^*| \end{aligned}$$

Desta dedução segue que:

$$\begin{aligned} \lim_{k\to\infty} |x_k - x^*| &\leqslant \lim_{k\to\infty} \gamma^k|x_0 - x^*| \\ &= |x_0 - x^*| \lim_{k\to\infty} \gamma^k \\ &= 0 \qquad\qquad [\text{Pois } \gamma < 1] \end{aligned}$$

Portanto, concluímos que $\lim_{k\to\infty} x_k = x^*$. ■

Considere a função $f(x) = x^3 - 5x + 3$. Um exemplo de processo iterativo é: $c(x) = \frac{1}{5}$ que nos leva a $g(x) = x + c(x).f(x) \Rightarrow g(x) = x + \frac{x^3-5x+3}{5} = \frac{x^3+3}{5}$.

3.5.1 Método da Iteração Linear

No método da iteração linear, dada uma função $f(x)$, utilizamos como função auxiliar $c(x) = 1$ que induz a função de iteração $g(x) = x + c(x).f(x) = x + f(x)$.

Dependendo da função de iteração, podemos obter um comportamento desejado (convergência para a solução) ou indesejado (divergência), conforme ilustração nas Figuras 3.12, 3.13, 3.14, e 3.15. Os pontos x^* que satisfazem a condição $x^* = g(x^*)$ são ditos pontos fixos de $g(x)$ e representam geometricamente os pontos de intersecção da reta $y = x$ com a curva $y = g(x)$.

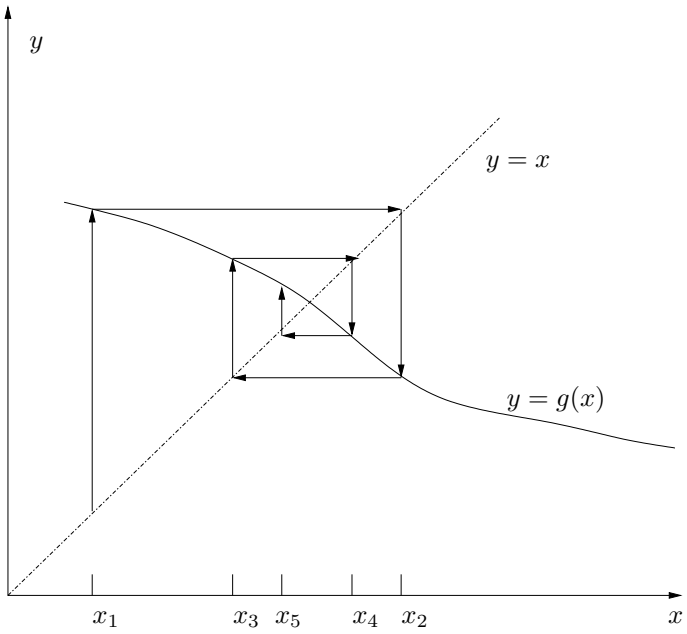


Figura 3.12: Convergência oscilante.

Exemplo 1

Para a função $f(x) = \frac{x^2}{10} - x + 1$, podemos obter uma função $g(x)$ fazendo $c(x) = 1$, que nos leva a $g(x) = \frac{x^2}{10} + 1$. Para o intervalo $I = [1, 1.5]$ vamos verificar se as condições do Teorema 3.2 são satisfeitas.

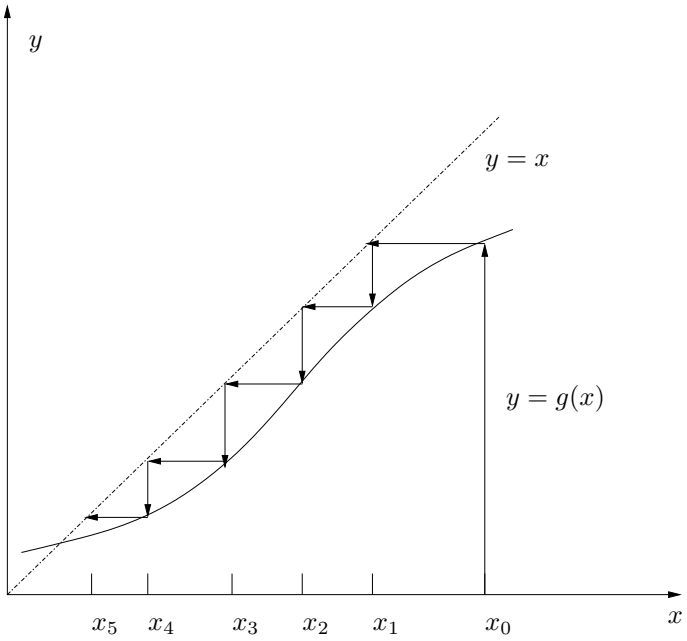


Figura 3.13: Convergência monotônica.

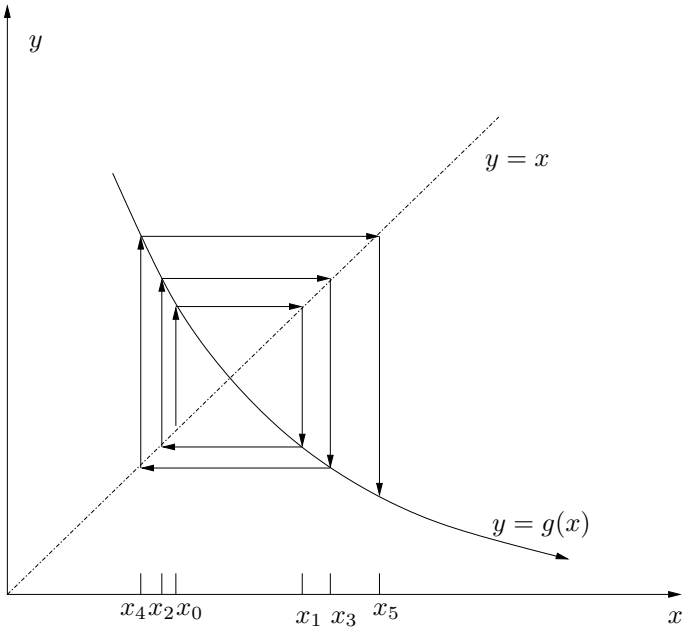


Figura 3.14: Divergência oscilante.

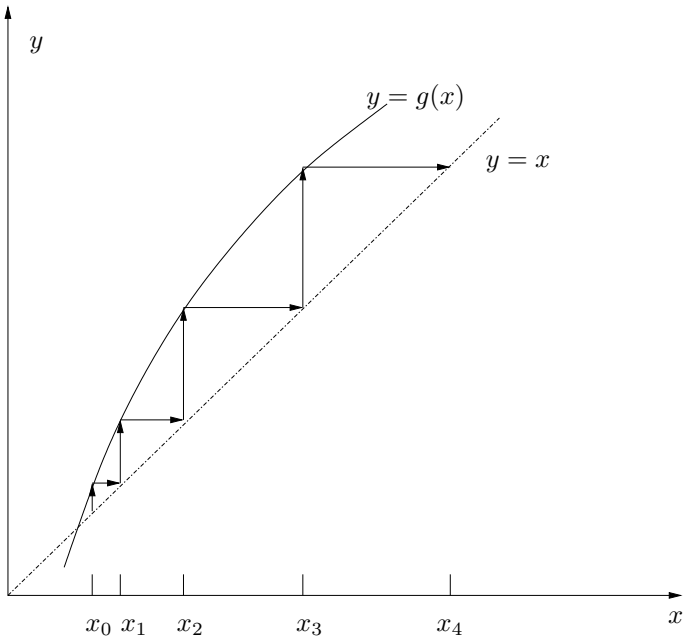


Figura 3.15: Divergência monotônica.

Primeiro, verificamos se $g(I) \subseteq I$. Seja $x \in I$, ou seja, $1 \leqslant x \leqslant 1.5$. Note que $g(x) \leqslant 1.5 \Leftrightarrow x^2 \leqslant 10 \times 0.5 \Leftrightarrow x \leqslant \sqrt{5} \Leftrightarrow x \leqslant 2.2361$. De maneira similar, $g(x) \geqslant 1 \Leftrightarrow x^2 \geqslant 0$. Portanto, se $x \in I$ temos que $g(x) \in I$.

Ainda resta verificar se $|g'(x)| < 1$. Note que $g'(x) = \frac{2x}{10}$. Para $x \in I$, $g'(x) \leqslant \frac{2\times 1.5}{10} = 0.3$.

Logo as condições do Teorema 3.2 são satisfeitas e, se tomarmos $x_0 \in I$, a seqüência $x_{k+1} = g(x_k), k = 0, 1, \ldots$ deve convergir para uma solução de $f(x) = 0$, conforme mostra o Teorema 3.3.

Para $x_0 = 1.5$ a seqüência de iterandos é descrita na Tabela 3.3 e ilustrada na Figura 3.16. Podemos observar que o método convergirá para o valor $x = 1.127\ 016\ 653\ 792\ 58$ que é uma das raízes exatas de $f(x) = 0$.

Exemplo 2

Para a função $f(x) = x - 2\sin(x)$, obtemos a função de iteração $g(x) = 2\sin(x)$ fazendo $c(x) = -1$. Vamos verificar se as condições do Teorema 3.2 são satisfeitas para o intervalo $I = (\frac{\pi}{2}, \frac{2\pi}{3}) = (90^0, 120^0) \approx (1.5708, 2.0994)$.

Observe que $0.866 \leqslant \sin(x) \leqslant 1$ para $x \in I$, portanto $g(x) \geqslant 2\sin(2\frac{\pi}{3}) = 1.7321$ para todo $x \in I$. Observe também que $g(x) \leqslant 2\sin(\frac{\pi}{2}) = 2$ para todo $x \in I$. Assim concluímos que $g(I) \subseteq I$.

Como $g'(x) = 2\cos(x)$, temos que $|g'(x)| \leqslant |2\cos(2\frac{\pi}{3})| = |2\times(-0.5)| = 1$.

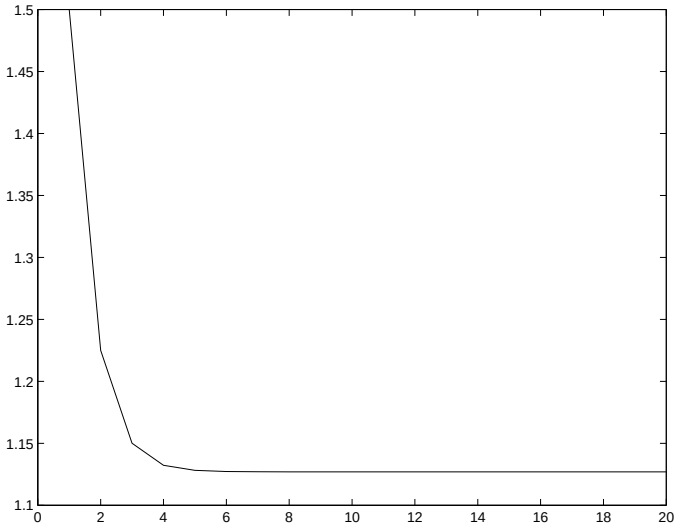


Figura 3.16: Exemplo de seqüência de iterandos de um processo iterativo.

Tabela 3.3: Seqüência de iterandos

k	x_k
0	1.5
1	1.225
2	1.150 0625
3	1.132 264 375 390 63
4	1.128 202 261 577 87
⋮	⋮
100	1.127 016 653 792 58

Assim, de acordo com os Teoremas 3.2 e 3.2, o processo iterativo convergirá para a raiz se $x_0 \in I$. A seqüência obtida para $x_0 = \pi/2$ é dada na Tabela 3.4.

3.5.2 Exercícios

i. Para a função $f(x) = \frac{x^2}{10} - x + 1$ do exemplo 1, foi demonstrado que o operador iterativo $g(x)$ obtido com $c(x) = 1$ é convergente dentro do intervalor $I = [1,\ 1.5]$. Calcule analiticamente a raiz $x^\star \in I$. Com $x_0 = 1.2$, aplique o processo iterativo até a iteração 4 e obtenha x_4. Verifique que o erro de truncamento observado $|x^\star - x_4|$ está dentro do limite de erro *a priori* estabelecido pelo Teorema 3.3. Repita a

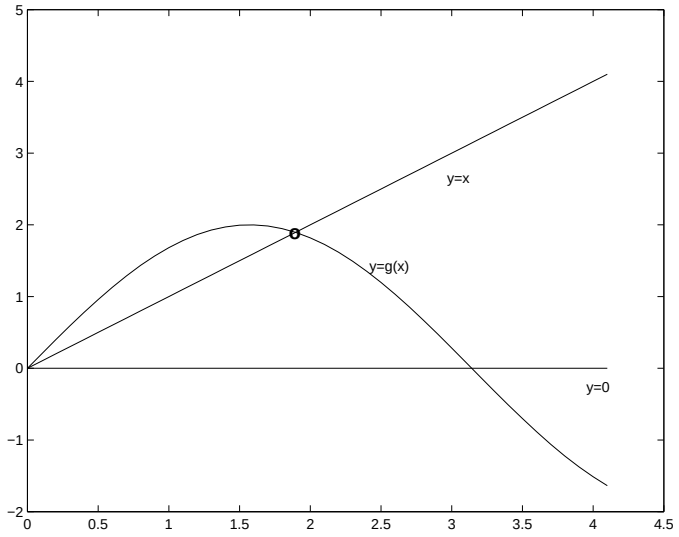


Figura 3.17: Método de ponto fixo para raiz de $f(x) = x - 2\sin(x)$.

verificação para o erro de truncamento *a posteriori*.

3.6 Método de Newton

Seja $f : \mathbb{R} \to \mathbb{R}$ uma função diferenciável. Dado x^k, o método de Newton utiliza a expansão de Taylor em torno do iterando x^k para obter o próximo iterando:

$$f(x^{k+1}) = f(x^k) + \frac{df(x^k)}{dx}(x^{k+1} - x^k) + O(\|x^{k+1} - x^k\|^2)$$

Para $\|x^{k+1} - x^k\|$ pequeno, fazemos

$$\begin{aligned} f(x^{k+1}) &\approx f(x^k) + \frac{df(x^k)}{dx}(x^{k+1} - x^k) \\ &= f(x^k) + f'(x^k)(x^{k+1} - x^k) \\ &= 0 \end{aligned}$$

que nos leva a relação

$$\begin{aligned} f'(x^k)(x^{k+1} - x^k) &= -f(x^k) \Rightarrow \\ x^{k+1} &= x^k - \frac{f(x^k)}{f'(x^k)} \end{aligned}$$

Tabela 3.4: Seqüência de iterandos

k	x_k
0	$\pi/2$
1	2
2	1.818
3	1.938
4	1.866
5	1.913
6	1.837
$\vdots$	$\vdots$
7	1.895 441 239

O comportamento do método de Newton pode ser ilustrado na Figura 3.18. O desenvolvimento acima pode ser condensado no algoritmo de Newton, cujos passos são dados abaixo.

Newton(f, x^0, ϵ)
1: $k \leftarrow 0$
2: **while** $|f(x^k)| > \epsilon$ **do**
3: $\quad x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}$
4: $\quad k \leftarrow k + 1$
5: **end while**
6: Saída x^k

3.6.1 Exemplo 1

Considere o polinômio $p(x) = x^3 - 5x^2 + 17x + 21$, cuja derivada é $p'(x) = 3x^2 - 10x + 17$. Tomando como ponto inicial $x^0 = -1$, o método de Newton gera a seqüência de iterandos dada conforme Tabela 3.5.

3.6.2 Exemplo 2

Para a função $f(x) = xe^{-x} - 0.2$, temos como derivada a função $f'(x) = e^{-x}(1-x)$. Tomando como ponto inicial $x^0 = 0$, o método de Newton produz a seqüência de iterandos listada na Tabela 3.6.

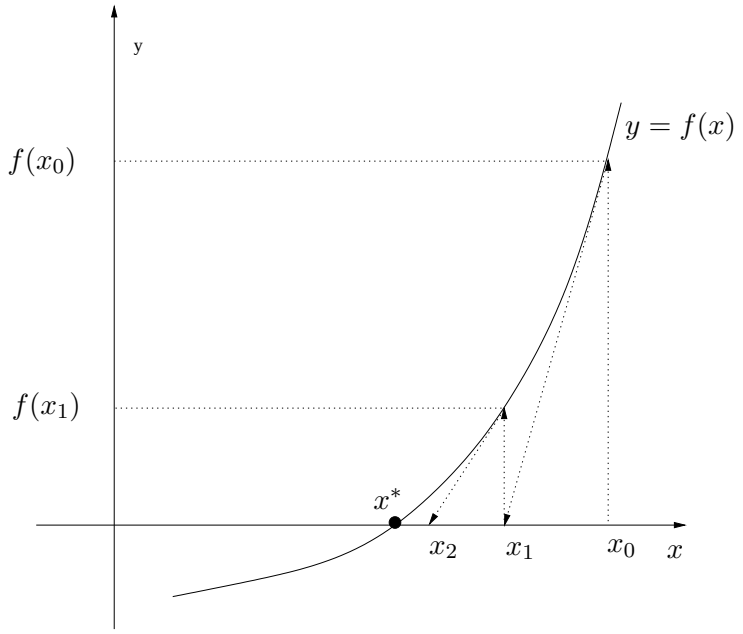


Figura 3.18: Ilustração do método de Newton

Tabela 3.5: Seqüência de iterandos produzida pelo método de Newton

k	x_k	$p(x^k)$	$p'(x^k)$
0	-1.0	-2	30
1	-0.933 333 333	-0.035 259 259	28.927 669
2	-0.932 115 256	-0.000 011 571	28.927 662
3	-0.932 114 856	-0.000 000 000	28.927 662
4	-0.932 114 856	-0.000 000 000	28.927 662

3.6.3 Detalhes do Método de Newton

Seja $f(x)$ uma função contínua e $p(x)$ uma aproximação de f, em torno do ponto x^0, obtida através da série de Taylor:

$$\begin{aligned} p(x) &= f(x^0) + f'(x^0)(x - x^0) + f''(x^0)\frac{(x - x^0)^2}{2!} + f'''(x^0)\frac{(x - x^0)^3}{3!} + \dots \\ &\quad + f^{(m)}(x^0)\frac{(x - x^0)^m}{m!} \end{aligned} \tag{3.6}$$

De acordo com o Teorema de Taylor, se a $(m+1)$-ésima derivada é contínua numa vizinhança de x^0, então o erro tem a forma:

$$f(x) - p(x) = (x - x^0)^{m+1}\frac{f^{(m+1)}(\varepsilon)}{(m+1)!}$$

Tabela 3.6: Seqüência de iterandos produzida pelo método de Newton

k	x_k	$f(x^k)$	$f'(x^k)$
0	0	-0.2	1.0
1	0.2	-0.036 253 849	0.654 984
2	0.255 350 689	-0.002 193 918	0.576 838
3	0.259 154 037	-0.000 009 554	0.571 713
4	0.259 171 101	-0.000 000 000	0.571 690

onde $\varepsilon = x^0 + \theta(x - x^0)$ para algum $\theta \in (0, 1)$. Na interpolação direta, a função f é interpolada num ponto próximo da raiz dessa função. Podemos supor, então, que a raiz desse polinômio dará uma boa aproximação para o zero da função.

O método de Newton leva em conta só a primeira derivada de (3.6). Logo, se o ponto aproximado é x^k, então

$$p(x^{k+1}) = f(x^k) + f'(x^k)(x^{k+1} - x^k)$$

Fazendo $p(x^{k+1}) = 0$, obtemos a nova aproximação da raiz:

$$\begin{aligned} f(x^k) + f'(x^k)(x^{k+1} - x^k) &= 0 \Rightarrow \\ x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)} & \end{aligned}$$

3.6.4 Convergência do Método de Newton

O método de Newton converge quadraticamente se $f'(x^*) \neq 0$ e se x^0 é suficientemente próximo de x^*.

Seja $I = [x^* - \delta, x^* + \delta]$ um intervalo em torno da solução x^* dentro do qual as seguintes condições são satisfeitas:

i. existe uma constante $m > 0$ tal que $|f'(x)| \geqslant m$ para todo $x \in I$; e

ii. existe uma constante $L > 0$ tal que $|f'(x) - f'(y)| \leqslant L|x - y|$ para todo $x, y \in I$.

Vamos mostrar que se $|x^k - x^*| \leqslant \delta$, então

$$|x^{k+1} - x^*| \leqslant \frac{L}{2m}|x^k - x^*|^2 \tag{3.7}$$

Em outras palavras, a desigualdade (3.7) é satisfeita com $c = L/(2m)$, garantindo dessa forma convergência quadrática para x^*. Denotando x^{k+1} por

x^+ e x^k por x, temos que

$$\begin{aligned} x^+ = x - \frac{f(x)}{f'(x)} \Rightarrow |x^+ - x^*| &= |x - \frac{f(x)}{f'(x)} - x^*| \\ &= \frac{|-f(x) - f'(x)(x^* - x)|}{|f'(x)|} \\ &= \frac{|f(x^*) - f(x) - f'(x)(x^* - x)|}{|f'(x)|} \end{aligned}$$

Lembre-se que $f(x^*) = 0$. Nós assumimos que $|f'(x)| \geqslant m$ e, portanto,

$$|x^+ - x^*| \leqslant \frac{|f(x^*) - f(x) - f'(x)(x^* - x)|}{m} \tag{3.8}$$

Para limitar o numerador, podemos proceder como segue

$$\begin{aligned} |f(x^*) - f(x) - f'(x)(x^* - x)| &= |\int_x^{x^*} (f'(u) - f'(x))du| \\ &\leqslant \int_x^{x^*} |f'(u) - f'(x)|du \\ &= L\int_x^{x^*} |u - x|du \\ &= L\frac{|x^* - x|^2}{2} \end{aligned} \tag{3.9}$$

Juntando (3.8) e (3.9) obtemos $|x^+ - x^*| \leqslant \frac{L}{2m}|x^* - x|^2$ o que prova (3.7). A desigualdade (3.7) por si só não garante que $|x^{k+1} - x^*| \to 0$. Entretanto, se assumirmos que $|x^0 - x^*| \leqslant \delta$ e $\delta \leqslant m/L$, então convergência quadrática segue imediatamente. Uma vez que $|x^0 - x^*| \leqslant \delta$, podemos aplicar (3.7) e obter:

$$|x^1 - x^*| \leqslant \frac{L}{2m}|x^0 - x^*|^2 \leqslant \frac{L}{2m}\delta^2 \leqslant \delta/2.$$

Uma vez que $|x^1 - x^*| \leqslant \delta$, podemos aplicar (3.7) e obter a desigualdade:

$$|x^2 - x^*| \leqslant \frac{L}{2m}|x^1 - x^*|^2 \leqslant \frac{L}{2m}\delta^2/4 \leqslant \delta/8.$$

Repetindo este processo, concluímos que:

$$|x^{k+1} - x^*| \leqslant \frac{L}{2m}|x^k - x^*|^2 \leqslant 2\left(\frac{1}{4}\right)^{2^k}\delta.$$

Observações:

1) Segundo o desenvolvimento acima, se $f(x^*) = 0$ e f' é uma função contínua, então é razoável assumir que as condições acima são satisfeitas para algum δ, m e L. Na prática, entretanto, é difícil encontrarmos valores para δ, m e L, portanto o resultado de convergência não nos oferece ajuda no sentido de escolhermos o ponto inicial x^0, mesmo porque isto demandaria conhecimento antecipado da existência de uma raiz.

2) O resultado nos diz que se x^0 é suficientemente próximo de x^*, então o método de Newton converge com uma taxa quadrática.

3.6.5 Problemas com o Método de Newton

Há situações, tipicamente dependendo do ponto inicial, em que o método de Newton não converge para uma solução. As Figuras 3.19 e 3.20 ilustram o caso onde o método de Newton pode entrar em laço infinito e o caso onde este pode divergir.

No caso da função $f(x) = x^3 - 2x + 2$, o método de Newton entra em laço infinito quando inicia a partir do ponto $x^{(0)} = 0$. O iterando seguinte será $x^{(1)} = 1$ e o seguinte ao seguinte será $x^{(2)} = 0$, portanto entra em laço infinito.

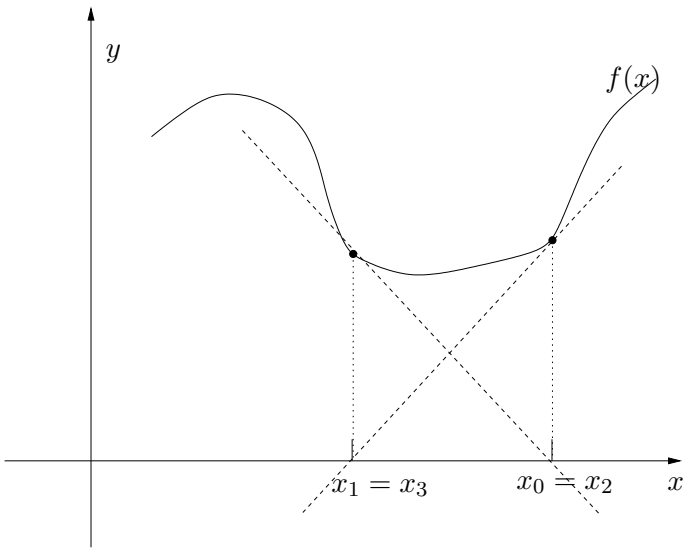


Figura 3.19: Exemplo onde o método de Newton entre em laço infinito.

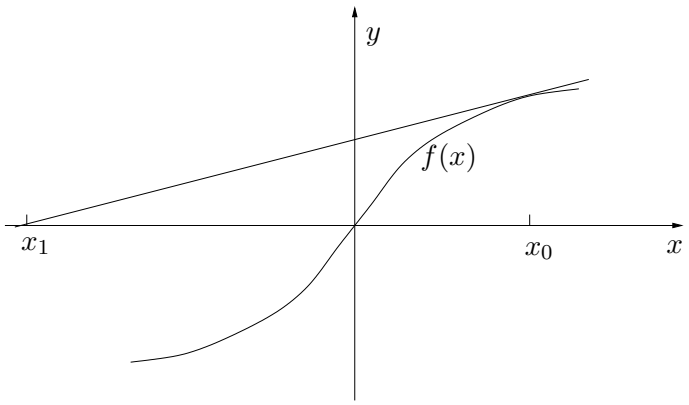


Figura 3.20: Exemplo onde ocorre divergência do método de Newton.

3.6.6 Exercícios

i. Para a função $f(x) = \frac{x^2}{10} - x + 1$, desenvolva o método de Newton obtendo o processo iterativo $g(x) = x - f(x)/f'(x)$. Para o intervalo $I = [1,\ 1.5]$, encontre o valor $m > 0$ tal que $f'(x) \geq m$ para todo $x \in I$. Encontre a constante Lipschitz $L > 0$ tal que $|f'(x) - f'(y)|/|x - y| \leq L$ para todo $x \in I$. Determine a raiz $x^\star \in I$. Encontre $\delta > 0$ que define o intervalo $\hat{I} = [x^\star - \delta, x^\star + \delta] \subseteq I$ tal que se o ponto inicial $x^0 \in \hat{I}$, o método de Newton converge quadraticamente para $x^\star$.

3.7 Métodos de Múltiplos Passos

Ao contrário dos métodos anteriores, os métodos de múltiplos passos usam os valores de f e suas derivadas, se for o caso, para vários pontos anteriores. O exemplo mais conhecido é o método das secantes.

3.7.1 Método das Secantes

Na prática, quando for muito complicado calcular derivadas e utilizar o método de Newton, podemos alternativamente usar o modelo linear tomando como base os dois valores mais recentes de f. A ideia do método pode ser vista na Figura 3.21 abaixo.

Partindo de duas aproximações x_0 e x_1, determinamos a reta que passa por $(x_0, f(x_0))$ e $(x_1, f(x_1))$. A intersecção dessa reta com o eixo x (definido pela equação $y = 0$) determina o próximo iterando x_2. Continuamos o

processo a partir de x_1 e x_2. Temos, portanto, o seguinte processo iterativo

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1})f(x_k)}{f(x_k) - f(x_{k-1})} \tag{3.10}$$

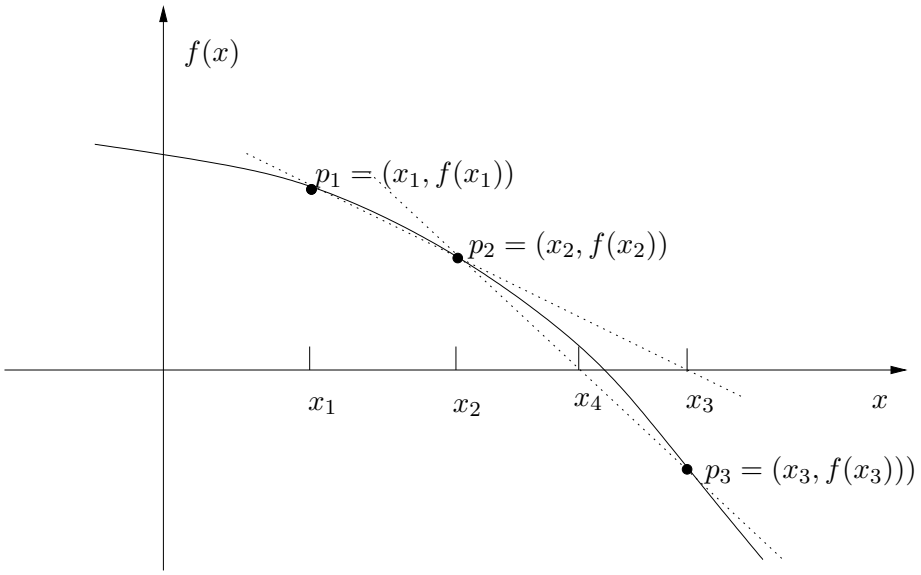


Figura 3.21: Ilustração do método das secantes.

Observações:

a) Neste método não necessitamos da característica que é fundamental no método da Falsa Posição, que exige que haja troca de sinal da função f no intervalo $[x_{k-1}, x_k]$.

b) Questões de *overflow/underflow* podem ocorrer. Tanto a fórmula (3.10) quanto a fórmula abaixo podem causar problemas de *cancelamento subtrativo* ou *overflow*:

$$x_{k+1} = \frac{x_k f(x_{k-1}) - x_{k-1} f(x_k)}{f(x_{k-1}) - f(x_k)} \tag{3.11}$$

Os problemas potenciais das fórmulas (3.10) e (3.11) podem ser evitados com a fórmula abaixo:

$$x_{k+1} = x_k - \left[(x_{k-1} - x_k)\frac{f(x_k)}{f(x_{k-1})}\right] / \left[1 - \frac{f(x_k)}{f(x_{k-1})}\right] \tag{3.12}$$

desde que $|f(x_k)| < |f(x_{k-1})|$. Se isto não ocorrer, então fazemos a troca de x_k por x_{k+1}.

Secantes$(x_0, x_1, f, \epsilon_1, \epsilon_2, L)$

$f_0 \leftarrow f(x_0); f_1 \leftarrow f(x_1)$
if $|f_1| > |f_0|$ **then**
  swap(x_0, x_1)
  swap(f_0, f_1)
end if
for $k = 0, 1, \ldots, L$ **do**
  if $|f_1| < \epsilon_1$ **then**
    Saída $(x_1, |f(x)| < \epsilon_1)$
    Pare
  end if
  $S \leftarrow f_1/f_0$
  $p \leftarrow (x_0 - x_1)s$
  $q \leftarrow 1 - s$
  $x_2 \leftarrow x_1 - p/q$
  if $|x_1 - x_2| < \epsilon_2|x_2|$ **then**
    Saída $(x_2, |x_1 - x_2| < \epsilon_2|x_2|)$
    Pare
  end if
  $f_2 \leftarrow f(x_2)$
  if $|f_2| > |f_1|$ **then**
    $x_0 \leftarrow x_2$
    $f_0 \leftarrow f_2$
  else
    $x_0 \leftarrow x_1$
    $f_0 \leftarrow f_1$
    $x_1 \leftarrow x_2$
    $f_1 \leftarrow f_2$
  end if
end for

Convergência do Método da Secante

Observa-se que o método é mais rápido que a iteração linear ou bissecção; contudo, pode ser mais lento que o método de Newton.

Teorema 3.5 *Se $f(x^*) = 0$ e $f'(x^*) \neq 0$ e f'' é contínua, então existe um intervalo aberto $N(x^*) = \{x : |x - x^*| < \epsilon\}$, $\epsilon > 0$, tal que se x_0 e x_1 estão em $N(x^*)$ e são distintos, então a seqüência dada pelo método da secante é tal que*

$$\lim_{k \to \infty} x_k = x^*$$

Exemplo

Para o problema de encontrar uma raiz da função $f(x) = x^4 - 14x^2 + 24x - 10$, vamos aplicar o método das secantes. O gráfico de $f(x)$ é ilustrado na Figura 3.9, indicando que existem quatro raízes. Aplicando o método das secantes com os pontos iniciais $x_0 = -5$ e $x_1 = 0$, obtemos as iterações dadas nas Figuras 3.22 e 3.23. O método converge para a raiz $x^* = 0.6705$.

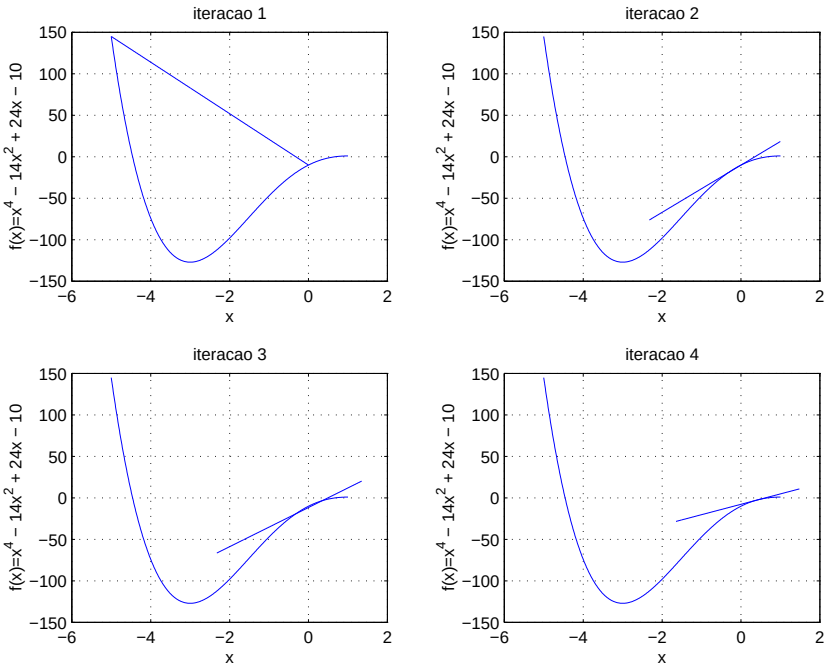


Figura 3.22: Iterações 1, 2, 3 e 4 do método das secantes

3.7.2 Breve Introdução à Interpolação Polinomial

Aqui faremos uma breve introdução ao problema de interpolação com enfoque em interpolação polinomial. A interpolação polinomial é base para o desenvolvimento do método de Müller.

Polinômio Interpolador de Lagrange

Dado um conjunto $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ de pontos de uma curva $f(x)$, sendo $y_k = f(x_k)$ para $k = 0, \ldots, n$, o polinômio interpolador de Lagrange é um polinômio $p_n(x)$ de grau n que passa por todos os pontos. A Figura 3.24 ilustra um polinômio $p_4(n)$ que atravessa os pontos indicados. O

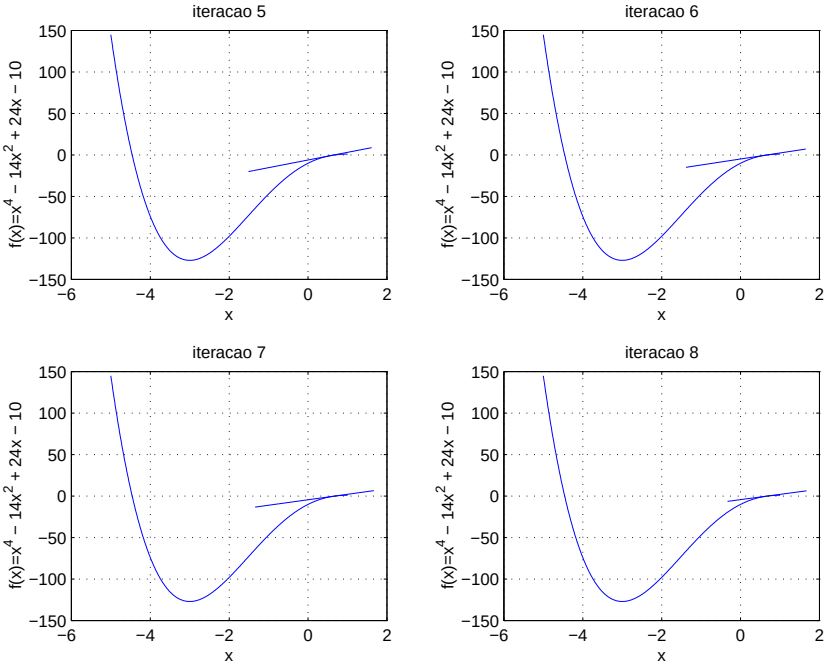


Figura 3.23: Iterações 5, 6, 7 e 8 do método das secantes

polinômio de Lagrange é definido por:

$$\begin{aligned} p_n(x) &= \sum_{k=0}^{n} L_k(x) \\ L_k(x) &= y_k \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \end{aligned} \tag{3.13}$$

De forma explícita, o polinômio (3.13) fica:

$$\begin{aligned} p_n(x) = & \frac{(x-x_1)(x-x_2)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)} y_0 \\ & + \frac{(x-x_0)(x-x_2)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)} y_1 \\ & + \quad \vdots \\ & + \frac{(x-x_0)(x-x_1)\dots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1})} y_n \end{aligned}$$

A fórmula acima foi inicialmente descoberta por Waring (1779), redescoberta por Euler em 1783, e publicada mais tarde por Lagrange em 1795 [4].

Para o caso de $n = 2$, o polinômio de Lagrange toma a forma:

$$p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2$$

Observe que $p_2(x)$ passa pelos pontos (x_0, y_0), (x_1, y_1) e (x_2, y_2).

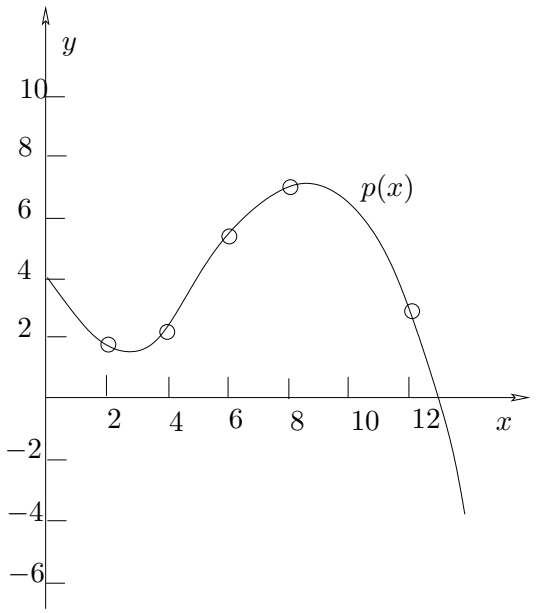


Figura 3.24: Ilustração de interpolação polinomial

Diferenças Divididas de Newton

O método das diferenças divididas de Newton é um maneira de encontrar um polinômio interpolador (um polinômio que passa por um conjunto particular de pontos). De forma similar ao polinômio interpolador de Lagrange para interpolação polinomial, o método de diferenças divididas de Newton encontra o polinômio único que atravessa os pontos.

O método de Newton usa a equação recursiva abaixo para realizar a tarefa:

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0} \tag{3.14}$$

A equação acima leva à fórmula de diferenças divididas de Newton para interpolação polinomial, sendo definida por:

$$\begin{aligned} p_n(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \ldots \\ &\quad + (x - x_0) \ldots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n] \end{aligned} \tag{3.15}$$

3.7.3 Método de Müller

No método da secante tomávamos dois pontos e por eles traçávamos uma reta (interpolação linear). Partiremos agora para a interpolação quadrática. Se tivermos três pontos distintos $(x_0, f(x_0))$, $(x_1, f(x_1))$ e $(x_2, f(x_2))$, podemos usar o método das diferenças divididas de Newton para computar o polinômio interpolador de 2^o grau:

$$p_2(x) = f(x_2) + (x - x_2)f[x_2, x_1] + (x - x_2)(x - x_1)f[x_2, x_1, x_0] \qquad (3.16)$$

sendo:

$$f[x_2, x_1] = \frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

$$f[x_1, x_0] = \frac{f(x_0) - f(x_1)}{x_0 - x_1}$$

$$f[x_2, x_1, x_0] = \frac{f[x_1, x_0] - f[x_2, x_1]}{x_0 - x_2}$$

Fazendo:

$$\begin{aligned} a &= f[x_2, x_1, x_0] \\ b &= f[x_2, x_1] + (x_2 - x_1)f[x_2, x_1, x_0] \\ c &= f(x_2) \end{aligned}$$

podemos escrever (3.16) de forma mais sintética:

$$p_2(x) = a(x - x_2)^2 + b(x - x_2) + c$$

A interpolação de uma curva $f(x)$ com um polinômio $p_2(x)$ é ilustrada na Figura 3.25. Em síntese, o método de Müller usa como próximo iterando o ponto x_3 que é raiz do polinômio $p_2(x)$, conforme indicado na figura. A equação:

$$p_2(x) = 0$$

tem duas soluções:

$$x - x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \implies x = x_2 + \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Dentre as duas soluções, tomamos aquela que é a mais próxima de x_2:

$$x_3 = x_2 + \frac{-b + \text{sign}(b)\sqrt{b^2 - 4ac}}{2a}$$

Müller(f, x_0, x_1, x_2)

1: $f_0 \leftarrow f(x_0)$
2: $f_1 \leftarrow f(x_1)$
3: $f_2 \leftarrow f(x_2)$
4: **for** $i = 1, 2, \ldots$ até satisfazer critério de parada **do**
5: $\quad f[x_2, x_1] \leftarrow \frac{f(x_1)-f(x_2)}{x_1-x_2}$
6: $\quad f[x_1, x_0] \leftarrow \frac{f(x_0)-f(x_1)}{x_0-x_1}$
7: $\quad f[x_2, x_1, x_0] \leftarrow \frac{f[x_1,x_0]-f[x_2,x_1]}{x_0-x_2}$
8: $\quad a \leftarrow f[x_2, x_1, x_0]$
9: $\quad b \leftarrow f[x_2, x_1] + (x_2 - x_1)f[x_2, x_1, x_0]$
10: $\quad c \leftarrow f(x_2)$
11: $\quad x_3 \leftarrow x_2 + \frac{-b+\text{sign}(b)\sqrt{b^2-4ac}}{2a}$
12: $\quad f_3 \leftarrow f(x_3)$
13: $\quad x_0 \leftarrow x_1,\ f_0 \leftarrow f_1$
14: $\quad x_1 \leftarrow x_2,\ f_1 \leftarrow f_2$
15: $\quad x_2 \leftarrow x_3,\ f_2 \leftarrow f_3$
16: **end for**
17: $\{x_2, f_2\}$

3.8 Aceleração de Aitken

A partir de dois métodos iterativos de mesma ordem é possível construir um novo método de ordem superior aos primitivos, conforme proposta de Aitken. Sejam dois métodos iterativos:

$$\begin{aligned} x_k^{(1)} &= \phi_1(x_{k-1}^{(1)}) \\ x_k^{(2)} &= \phi_2(x_{k-1}^{(2)}) \end{aligned}$$

ambos de ordem de convergência p, que convergem para $x = \alpha$. Podemos então construir uma função $\phi(x)$ dada por

$$\phi(x) = \frac{x\phi_1[\phi_2(x)] - \phi_1(x)\phi_2(x)}{x - \phi_1(x) - \phi_2(x) + \phi_1[\phi_2(x)]}$$

Então o método iterativo $x_k = \phi(x_{k-1})$, $k = 1, 2, \ldots$ tem ordem de convergência superior a p, desde que seja satisfeita a condição:

$$(\phi_1(\alpha) - 1)(\phi_2(\alpha) - 1) \neq 0$$

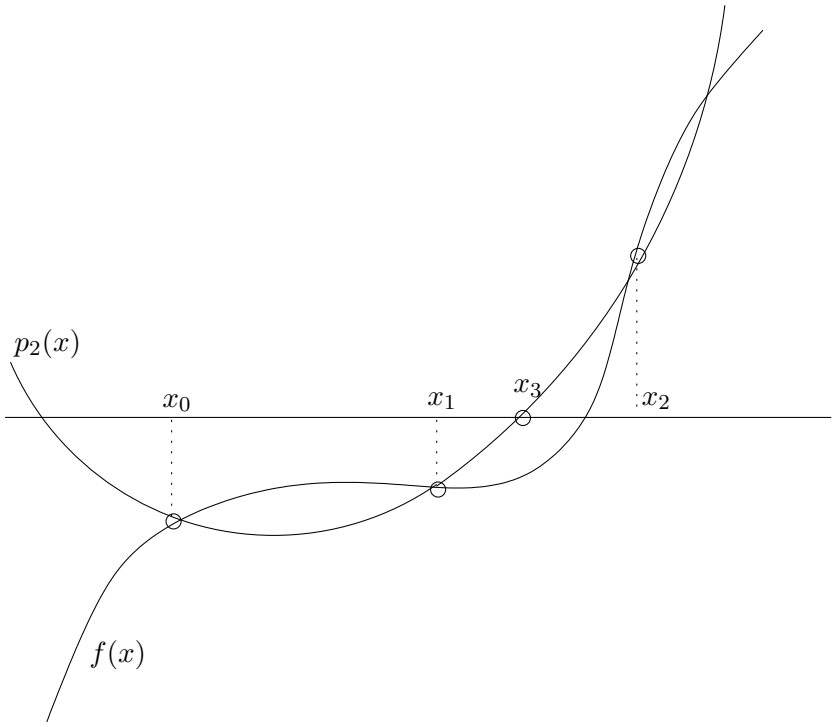


Figura 3.25: Interpolação da curva $f(x)$ com um polinômio de 2^o grau $p_2(x)$ que atravessa os pontos $(x_0, f(x_0))$, $(x_1, f(x_1))$ e $(x_2, f(x_2))$

3.9 Exercícios

Exercício 3.1 Prof. Wellis afirma que se o algoritmo da bisecção pode ser aplicado para encontrar a raiz de uma função f, então também podemos aplicar o método de Newton. Você concorda ou discorda da afirmação?

Exercício 3.2 Seja f uma função definida por:

$$f(x) = \begin{cases} \sqrt{x} & se\ x \geqslant 0 \\ -\sqrt{-x} & se\ x < 0 \end{cases}$$

Vamos aplicar o método de Newton para encontrar uma raiz de $f(x)$. Responda as questões abaixo.

a) Existe uma região de convergência?

b) Existe uma região de divergência?

c) Existe uma região onde o método entra em laço infinito?

d) Para os itens que você respondeu afirmativamente, dê as respectivas regiões.

Exercício 3.3 Seja $g(x) = x + c(x)f(x)$ um processo iterativo, obtido pelo método de ponto fixo com iteração linear, para encontrar uma raiz da função $f(x)$. Prof. Wellis afirma que se x^* é tal que $f(x^*) = 0$, então $g'(x^*) = 0$. Você concorda ou discorda da afirmação?

Exercício 3.4 Seja $f(x) = x^2 - 2xe^{-x} + e^{-2x}$ uma função. Aplique o método de Newton para encontrar uma aproximação de uma raiz x^* de f, $f(x^*) = 0$. A aproximação x encontrada deve ser tal que $|f(x)| < 10^{-3}$. Execute no máximo 10 iterações do método.

Exercício 3.5 Uma bóia esférica de raio R e densidade ρ, ao flutuar na água, afunda de um quantidade x, dada por $x^3 + 2Rx^2 - 4\rho R^3 = 0$. Encontre o afundamento quando $R = 3$ e o material é cortiça ($\rho = 0.25$), usando o método da bisecção.

Exercício 3.6 Considere o sistema de equações abaixo:

$$\begin{cases} y^2 - 3x^2 - 5x - 5 &= 0 \\ y + x^2 - x &= 0 \end{cases}$$

Encontre uma solução $z^* = (x^*, y^*)$ para o sistema acima usando um método numérico. Sabemos que existe uma solução para $-1.5 \leqslant x^* \leqslant 0$.

Exercício 3.7 Seja $f(x) = xe^{-x}$ e $c(x) = -\frac{1}{e^{-x}(1+x)}$. Prof. Martins afirma que para qualquer $x_0 \in I = [0, 1)$ o operador de ponto fixo $g(x)$ resultante produz uma série $x_0, x_1, x_2, \ldots$ tal que $\lim_{k\to\infty} x_k = x^*$ sendo $f(x^*) = 0$. Se você concorda com a afirmação, apresente uma justificativa formal. Caso contrário, dê um contra-exemplo.

Exercício 3.8 Considere a função raiz quadrada $f(x) = \sqrt{x}$. É possível aplicar o método de Newton (em uma ou mais variáveis) para calcular $f(x)$? Se sim, mostre como o método de Newton pode ser aplicado para calcular $\sqrt{x}$? (*No seu computador/liguagem de programação, a função* $\sqrt{x}$ *não está disponível*)

Exercício 3.9 Desejamos encontrar uma raiz da função $f(x) = x^4 - 14x^2 + 24x - 10$. Sabemos que existe uma raiz no intervalo $I = [-5, 0]$.

Tarefas e observações:

- Traçe o gráfico da função no intervalo $[-5, 5]$ e no intervalo $[-5, 0]$.
- Encontre uma raiz em I aplicando os métodos abaixo:
  - O método da bisecção
  - O método da falsa posição
  - O método de Newton com $x^0 = 0$ e com $x^0 = -5$
  - O método das secantes com $x_0 = -5$ e $x_1 = 0$
- Para cada um dos métodos acima, desenhe a curva da solução aproximada em função do número de iteração.
- As implementações devem ser realizadas em Matlab, Octave ou Scilab. Apresentar o código fonte juntamente com os resultados.

Exercício 3.10 Para a função F dada por:

$$F(x, y) = \begin{bmatrix} (x - y + 1/4)e^{-x^2-y^3} \\ 2x^2 + \cos y \end{bmatrix}$$

calcule o Jacobiano ∇F. Obtenha o valor do Jacobiano no ponto $(x^0, y^0) = (1, 0)$. Obtenha o iterando (x^1, y^1) aplicando uma iteração do método de Newton a partir do ponto (x^0, y^0).

Exercício 3.11 Responda as questões abaixo.

i. Seja f uma função contínua em $I = [a, b]$, tal que $f(a).f(b) < 0$. Podemos aplicar o método da bisecção e o método de Newton para encontrar tal raiz?

ii. Seja $f : \mathbb{R} \to \mathbb{R}$ uma função contínua e diferenciável no intervalo $I = [a, b]$. Sabemos que existe exatamente um $x^* \in I$ tal que $f(x^*) = 0$. Sabemos que $f(a).f(b) < 0$. Sabemos que o método de Newton converge para $x^* \in I$. Seja $x^0_{bs}, x^1_{bs}, \ldots$ a sequência de pontos gerada pelo método da bisecção. Seja $x^0_{Nw}, x^1_{Nw}, \ldots$ a sequência de pontos gerada pelo método de Newton. Se $x^0_{Nw} = x^0_{bs} \neq x^*$, podemos afirmar que para algum i temos que $x^i_{Nw} = x^*$ e $x^i_{bs} \neq x^*$ visto que o Método de Newton tem convergência quadrática, enquanto o método da bisecção tem convergência linear?

iii. Seja $f : \mathbb{R} \to \mathbb{R}$ uma função qualquer para qual sabemos que existe uma raiz *única* x^* no intervalo $[a, b]$. Assumindo que os métodos da bisecção e da falsa posição convergem para x^*, podemos afirmar que o método da falsa posição converge mais rapidamente para o ponto x^* do que o método da bisecção?

iv. Seja $g : \mathbb{R} \to \mathbb{R}$ um operador de ponto fixo definido por $g(x) = x + c(x)f(x)$. Se $f(x) = ax^3 + bx^2 + cx + d$ e $c(x) = 1$, então $g(x)$ é um polinômio de grau 0, 1, 2 ou 3. Podemos afirmar que $g(x)$ terá nenhum, um, dois ou no máximo três pontos fixos?

v. Uma vez que o método de Newton exige que a função $f : \mathbb{R} \to \mathbb{R}$ seja contínua e diferenciável, enquanto que o método da bisecção exige que a função seja apenas contínua, podemos afirmar que o método da bisecção é mais geral? Em outras palavras, se podemos encontrar a raiz de uma função usando o método de Newton, então também podemos utilizar o método da bisecção?

Exercício 3.12 Considere as funções abaixo:

$$\begin{aligned} g(x, y) &= 5.3e^x - x - y \\ h(x, y) &= x^3 + 2.21x + 2y \end{aligned}$$

Desejamos encontrar $x, y \in \mathbb{R}$ tal que $g(x, y) = 0$ e $h(x, y) = 0$. É possível resolver este problema resolvendo o problema de encontrar uma raiz de uma função $f(x)$, $f : \mathbb{R} \to \mathbb{R}$?

Exercício 3.13 Obtenha o polinômio característico $p(\lambda)$ da matriz A abaixo e gere o gráfico $p(\lambda) \times \lambda$. Calcule os autovalores de A aplicando o método da falsa posição.

$$A = \begin{bmatrix} 2 & 0 & 4 & 5 & 0 \\ 0 & 0 & 0 & 5 & -1 \\ 4 & 0 & 0 & 0 & 1 \\ 5 & 5 & 0 & 2 & 0 \\ 0 & -1 & 1 & 0 & -2 \end{bmatrix}$$

Exercício 3.14 Seja $f(x) = x^4 - 2xe^{-x} + e^{-2x} - 4 - x$ uma função. Aplique o método de Newton para encontrar uma aproximação de uma raiz x^* de f, ou seja, x^* tal que $f(x^*) = 0$. A aproximação x encontrada deve ser tal que $|f(x)| < 10^{-4}$. Execute no máximo 20 iterações do método.

Exercício 3.15 Uma porta com formato descrito na Figura 3.26 é composta de uma parte retangular e uma parte semi-circular. A área da parte retangular da porta deve ser maior ou igual a $3.8m^2$. Encontre as dimensões x e y tal que o perímetro seja minimizado. Modele o problema matematicamente e encontre as dimensões desejadas utilizando o método da bisecção.

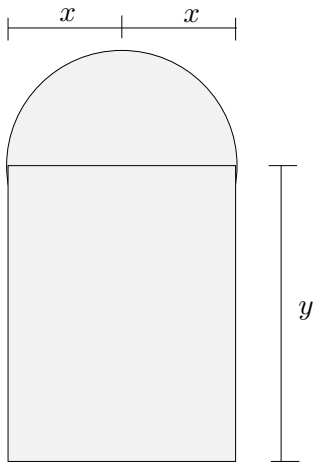


Figura 3.26: Porta semi-circular

Exercício 3.16 Seja $x_k = \frac{1}{k}$ uma sequência para $k = 1, 2, \ldots, \infty$. Questões:

i. Podemos afirmar que $\{x_k\}_{k=1}^{\infty}$ é convergente? Justifique a resposta com base na Definição 3.1.

ii. Se a seqüência for convergente, qual é a taxa de convergência? Justifique a resposta com base na Definição 3.2.

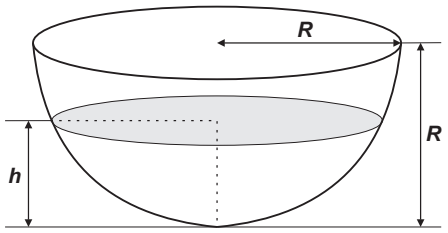


Figura 3.27: Reservatório tipo semiesfera.

Exercício 3.17 Seja $g(y) = (y^2 + x)/2y$ um processo iterativo onde $x > 1$ é uma constante conhecida. Seja $I = (1, x)$ o intervalo de incerteza. Questões:

i. Podemos afirmar que existe um ponto fixo $y^* \in I$?

ii. Caso exista ponto fixo, podemos afirmar que ele é único?

iii. Se existir ponto fixo y^* único, podemos afirmar que com $y_0 \in I$ o processo iterativo $y_{k+1} = g(y_k), k = 0, 1, \ldots$ converge para y^*?

Exercício 3.18 Um reservatório na forma de semiesfera com raio $R = 4m$ foi instalado em um prédio recém-construído. Ocorreu um erro de dimensionamento do reservatório: o volume de água total do reservatório é bem maior que o limite de $50m^3$ estabelecido em projeto hidro-sanitário. Assim, é necessário determinar o nível h máximo que a água pode atingir no reservatório para não ultrapassar o limite volumétrico. O reservatório é ilustrado na Figura 3.27. O volume de uma calota esférica é dado por:

$$V = \frac{\pi}{3} h^2 (3R - h)$$

sendo R o raio e h a altura.

Determine a altura h com precisão 10^{-6} utilizando o método do ponto fixo com $c(x) = -\frac{1}{16Rx}$. Mostre que o processo iterativo resultante converge dentro do intervalo $I = [2,\ 4]$ segundo a teoria do ponto fixo.

Capítulo 4

Resolução de Sistemas de Equações Lineares

Capítulo

4.1 Revisão

Sistemas de equações lineares surgem na solução de vários problemas teóricos e práticos. Dentre eles, ressaltamos:

1) Inteligência artificial: resolução das equações de Bellman

2) Teoria dos grafos: fluxo factível

3) Circuitos elétricos: análise de sistemas lineares (forma real e complexa)

4) Teoria de controle: sistemas lineares e a linearização de sistemas não lineares

O problema de resolver um sistema de equações lineares pode ser colocado como segue: dado uma matriz $A \in \mathbb{R}^{m\times n}$ e um vetor $b \in \mathbb{R}^{m\times 1}$, encontre $x \in \mathbb{R}^{n\times 1}$ tal que $Ax = b$.

A matriz de coeficientes e os vetores podem ser expressos como segue:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

onde $a_{ij} \in \mathbb{R}, i = 1, \ldots, m, j = 1, \ldots, n$ e

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix},$$

onde $b_j \in \mathbb{R}, j = 1, \ldots, m$

Definição 4.1 *$A \in \mathbb{R}^{m \times n}$ é dita quadrada se $m = n$*

Definição 4.2 *$A \in \mathbb{R}^{n \times n}$ é dita diagonal se $A = diag(d_1, d_2, \ldots, d_n)$, ou seja,*

$$A = \begin{bmatrix} d_1 & 0 & \ldots & 0 \\ 0 & d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{bmatrix}$$

Definição 4.3 *A matriz $A \in \mathbb{R}^{n \times n}$ abaixo é dita matriz triangular inferior se:*

$$A = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$

em outras palavras,

$$A = [a_{ij}] \text{ e } a_{ij} = 0 \text{ para todo } a_{ij} \text{ onde } j > i.$$

Definição 4.4 *Se $A \in \mathbb{R}^{n \times n}$, então A possui um determinante $det(A)$ definido como:*

$$det(A) = a_{11} det(M_{11}) - a_{12} det(M_{12}) + a_{13} det(M_{13}) - \ldots$$

onde M_{1k} é a matriz obtida de A eliminando-se a linha 1 e a coluna k. O determinante de uma matriz $A \in \mathbb{R}^{1 \times 1}$ é a própria matriz.

Definição 4.5 *$A \in \mathbb{R}^{n \times n}$ é dita não-singular se $det(A) \neq 0$.*

Definição 4.6 *Se $A \in \mathbb{R}^{n \times n}$ é uma matriz triangular (inferior ou superior) então:*

$$det(A) = a_{11} a_{22} \ldots a_{nn}$$

Definição 4.7 *Se $A \in \mathbb{R}^{n\times n}$ é não-singular então existe $A^{-1} \in \mathbb{R}^{n\times n}$ tal que $AA^{-1} = I$, onde $I = diag(1, 1, \ldots, 1)$*

Definição 4.8 *A matriz adjunta de A, denotada por $Adj(A)$, é definida por:*

$$Adj(A) = \begin{bmatrix} C_{11}(A) & \cdots & C_{n1}(A) \\ \vdots & \ddots & \vdots \\ C_{1n}(A) & \cdots & C_{nn}(A) \end{bmatrix}$$

onde $C_{ij}(A)$ é o (i,j)-ésimo cofator de A, ou seja, $C_{ij}(A) = (-1)^{i+j} det(M_{ij}(A))$ sendo $M_{ij}(A)$ a matriz obtida ao se remover a linha i e coluna j de A.

Teorema 4.1 *Dada uma matriz não-singular $A \in \mathbb{R}^{n\times n}$, a sua inversa pode ser obtida a partir da matriz adjunta:*

$$A^{-1} = \frac{1}{det(A)} Adj(A)$$

Teorema 4.2 *A inversão apresenta as seguintes propriedades:*

i. *A inversa de A^{-1} é A, ou seja, $(A^{-1})^{-1} = A$.*

ii. *A inversa da transposta é a transposta da inversa: $(A^T)^{-1} = (A^{-1})^T$.*

iii. $(AB)^{-1} = B^{-1}A^{-1}$.

iv. $det(A^{-1}) = \frac{1}{\det(A)}$.

v. *A inversa de uma matriz triangular inferior (superior) é também triangular inferior (superior).*

Definição 4.9 *Seja $A \in \mathbb{R}^{m\times n}$ uma matriz dada por*

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

então a tranposta de A, denotada por A^T é dada por:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \ldots & a_{mn} \end{bmatrix}$$

Definição 4.10 *Se $A = A^T$, então A é dita uma matriz simétrica.*

Definição 4.11 *Uma matriz $A \in \mathbb{R}^{n\times n}$ é dita positiva definida se, e somente se, $x^T Ax > 0$ para todo $x \neq 0$. É dita positiva semi-definida se $x^T Ax \geq 0$ para todo x.*

Definição 4.12 *Seja $A \in \mathbb{R}^{n\times n}$ uma matriz quadrada. Se $Ax = \lambda x$ para $\lambda \in \mathbb{C}$ e $x \in \mathbb{R}^n$, então λ é um autovalor e x é um autovetor da matriz A.*

Teorema 4.3 *Se $A = A^T$, então todos os autovalores de A são números reais.*

Teorema 4.4 *Uma matriz $A = A^T$ é positiva definida (semi-definida) se, e somente se, os autovalores de A são todos positivos (não negativos).*

Definição 4.13 *Dada uma matriz $A \in \mathbb{R}^{m\times n}$, temos que:*

1) $range(A) = \{y \in \mathbb{R}^m : y = Ax$, para algum $x \in \mathbb{R}^n\}$ é o espaço gerado por A;

2) $null(A) = \{x \in \mathbb{R}^n : Ax = 0\}$; e

3) $rank(M)$ denota o posto da matriz M, i.e., o número máximo de colunas linearmente independentes de M;

Podemos examinar uma matriz $A \in \mathbb{R}^{m\times n}$ como uma geradora para a função linear $f : \mathbb{R}^n \to \mathbb{R}^m$ onde $f(x) = Ax$ para $x \in \mathbb{R}^n$. Assim, o espaço gerado por A, $range(A)$, é o conjunto das imagens da função $f(x)$. O espaço nulo de A, $null(A)$, é o conjunto dos valores que tem como imagem o vetor nulo. A Figura 4.1 ilustra estes conceitos.

Definição 4.14 *Seja S um espaço vetorial. O número máximo de vetores linearmente independentes de S é dito dimensão de S e denotado por $dim(S)$.*

Teorema 4.5 *Considere o sistema de equações lineares*

$$Ax = b \tag{4.1}$$

onde $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, e $b \in \mathbb{R}^m$. Então o sistema (4.1):

- *tem solução se $b \in range(A) \Leftrightarrow rank(A) = rank([A|b])$;*
- *não tem solução se $b \notin range(A) \Leftrightarrow rank(A) < rank([A|b])$;*
- *tem solução única se $rank(A) = rank([A|b]) = n$; e*

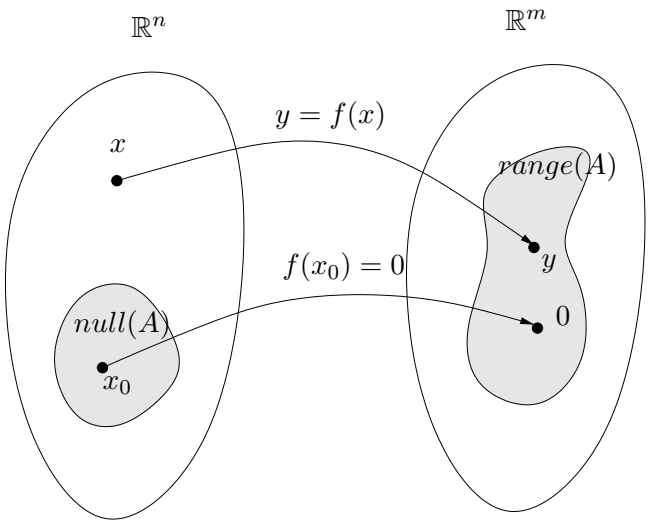


Figura 4.1: Ilustração dos conceitos de espaço gerado e espaço nulo de uma matriz A

- *tem um número infinito de soluções se* $rank(A) = rank([A|b]) < n$.

Teorema 4.6 *Para o sistema de equações lineares (4.1), vale*

$$dim(range(A)) + dim(null(A)) = n$$

Como exemplo, considere o sistema de equações lineares dado por:

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

Podemos verificar que $rank(A) = 2$ (note que $rank(A) \leqslant \min\{m, n\} = 2$). Portanto, $rank(A) = rank([A|b])$ o que nos leva a concluir que o sistema tem solução. Uma vez que $rank(A) < n$, o sistema tem infinitas soluções. A partir do Teorema 4.6, podemos verificar também que $dim(range(A)) = rank(A) = 2$. Portanto, $n = 3 \Rightarrow dim(null(A)) = n - dim(range(A)) = 1$ o que prova que o espaço nulo tem elementos não nulos.

Definição 4.15 *Uma norma (vetorial) é uma função* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ *satisfazendo:*

a) $x \in \mathbb{R}^n$, $\|x\| \geqslant 0$ *e* $\|x\| = 0 \Leftrightarrow x = 0$

b) $x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha|\|x\|$

c) $x, y \in \mathbb{R}^n$, $\|x + y\| \leqslant \|x\| + \|y\|$

Exemplos

Três exemplos de normas vetorais frequentemente encontradas são:

Norma Euclidiana: $\|x\|_2 = \sqrt{\sum_{j=1}^{n} (x_j)^2}$

Norma da soma: $\|x\|_1 = \sum_{j=1}^{n} |x_j|$

Norma do máximo: $\|x\|_\infty = \max |x_j| : j = 1, \ldots, n$

onde assumimos que $x \in \mathbb{R}^n$.

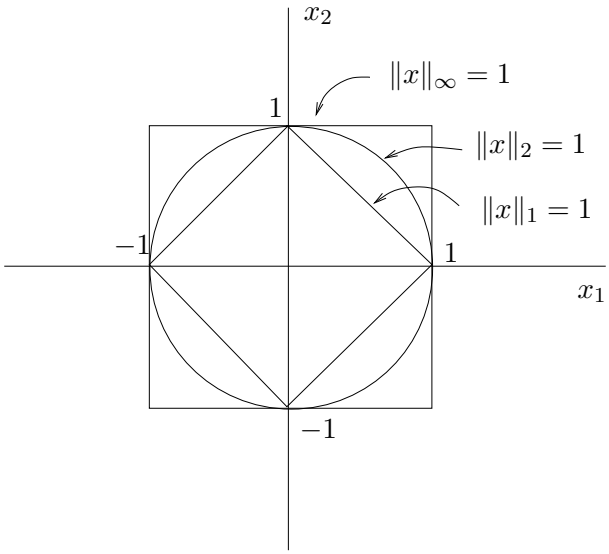


Figura 4.2: Ilustração das normas $\|\cdot\|_\infty, \|\cdot\|_2$ e $\|\cdot\|_1$.

Definição 4.16 *Uma norma matricial é uma função* $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ *satisfazendo as propriedades:*

a) $A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$ *e* $\|A\| = 0 \Leftrightarrow A = 0$

b) $A \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha|\|A\|$

c) $A, B \in \mathbb{R}^{n\times n}$, $\|A + B\| \leqslant \|A\| + \|B\|$

d) $A, B \in \mathbb{R}^{n\times n}$, $\|AB\| \leqslant \|A\| \cdot \|B\|$

Definição 4.17 *Toda norma vetorial* $\|\cdot\|$ *induz uma norma matricial* $\|\cdot\|$ *dada por:*

$$\|\cdot\| \quad = \quad \max_{x \neq 0} \quad \frac{\|Ax\|}{\|x\|},$$

onde $A \in \mathbb{R}^{m\times n}$.

Vamos denotar $gain(x) = \frac{\|Ax\|}{\|x\|}$, então:

$$\|A\| = \max_{x \neq 0} gain(x).$$

$gain(x)$ é o fator de ampliação do operador $f(x) = Ax$ na direção de x. Certamente, o ganho geralmente depende de x. $gain(x)$ pode ser grande para certos x e pode ser pequeno para outros. Se $\|A\| < 1$, então $\|Ax\| < \|x\|$ para qualquer $x \in \mathbb{R}^n$. Se $\|A\|$ é grande, então existe um x tal que o operador $f(x)$ amplifica, ou seja $\|Ax\| > \|x\|$.

Exemplo

Para $\|\cdot\| = \|\cdot\|_2$, considere a norma matricial, induzida por $\|\cdot\|$. Então $\|I$
é dada por:

$$\|I\| = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x \neq 0} \frac{\|x\|}{\|x\|} = 1$$

Exercício

Calcule, usando a definição de ganho, $\|A\|$ sendo A dada por:

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

Respostas: $\|A\|_1 = 1$, $\|A\|_2 = 1$, $\|A\|_\infty = 1$.

Algumas normas frequentemente utilizadas são as seguintes:

1) *Norma máximo das colunas:*

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \left(\sum_{i=1}^{n} |a_{ij}|\right)$$

2) *Norma máximo das linhas:*

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant n} \left(\sum_{j=1}^{n} |a_{ij}|\right)$$

3) *Norma Euclidiana:*

$$\|A\|_2 \;=\; \left(\sum_{i=1}^{n}\sum_{j=1}^{n} {a_{ij}}^2\right)^{\frac{1}{2}}$$

Propriedades:

a) $\|Ax\|_\infty \leqslant \|A\|_\infty \|x\|_\infty$

b) $\|Ax\|_1 \leqslant \|A\|_1 \|x\|_1$

c) $\|Ax\|_2 \leqslant \|A\|_2 \|x\|_2$

Aritmética matricial:

i) $C = A + B = B + A$, $A, B \in \mathbb{R}^{m\times n}$ e $c_{ij} = a_{ij} + b_{ij}$

ii) $C = AB$, $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$ e $c_{ij} = \sum_{t=1}^{k} a_{it} b_{tj}$

4.2 Erros Computacionais na Solução de $Ax = b$

Dois tipos de algoritmos para resolver sistemas do tipo $Ax = y$, onde $A \in \mathbb{R}^{m\times n}$ (ou $A \in \mathbb{C}^{m\times n}$)

4.2.1 Tipos de Algoritmos

A) Métodos Diretos

Um método é dito direto quando este utiliza um número finito de passos. Exemplos: Método de Cramer e Gauss Cramer.

1) Cramer: $x_j = \frac{det(A_j)}{det(A)}$ quando $A \in \mathbb{R}^{n\times n}$

2) Gauss: pivoteamento

3) Decomposição LU: Escrever $A = LU$, onde L é triangular inferior e U é triangular superior.

$$LUx = b \rightarrow Lz = b \text{ e } z = Ux$$

4) Decomposição Cholesky: Se $A = A^T$ e $A > 0$ (positiva definida) então podemos escrever A como produto dos fatores Cholesky, i.e., $LL^T = A$ para L triangular inferior.

5) Método do Gradiente Conjugado (Otimização)

B) Métodos Iterativos

Os métodos de Jacobi e Gauss-Seidel consistem de processos iterativos da forma:

$$x_{k+1} = G(x_k)$$

4.2.2 Tipos de Erros Computacionais nos Algoritmos

A) Algoritmos Diretos

Dado o sistema $Ax = y$, este será representado em uma máquina de ponto flutuante F, $F = (b, n, e_1, e_2)$

$$Ax = y \rightarrow \begin{cases} \square A \square x = \square y \\ \square A \text{ é a representação de } A \text{ em } F \\ \square x \text{ é a representação de } x \\ \square y \text{ é a representação de } y \end{cases} .$$

O algoritmo que executa as operações em F vai gerar um erro, obtendo uma solução $\square x$ que poderá ser diferente de x.

$\square' x = \square\square x$ é o erro total do método que consiste em:

$$E_{TOTAL} = (\text{Erro de entrada}) + (\text{Erro de aritmética})$$

O erro de aritmética é dado ao erro de arredondamento:

$$E_{TOTAL} = (x - \square x) + (\square x - \square' x) = E_A$$

B) Algoritmos Iterativos

Da mesma forma que os algoritmos diretos, os algoritmos iterativos cometem erros de entrada e de aritmética. Métodos iterativos também introduzem erros de discretização, ou seja, trunca-se uma sequência infinita $\{x_k : k = 0, 1, \ldots\}$ por uma sequência finita.

4.3 Etapas da Solução de Sistemas Lineares

Abaixo descrevemos os três passos básicos para solução de um sistema de equações lineares.

1) Primeira Etapa: Descomplexificação

Transformar um sistema complexo $Ax = y$ em um sistema real $\tilde{A}x = \tilde{y}$ equivalente.

2) Segunda Etapa: Estruturação

Escolher um algoritmo eficiente para calcularmos a solução, levando em consideração a estrutura de A, seu tamanho, esparsidade e simetria. Uma matriz $A \in \mathbb{R}^{m\times n}$ é dita esparsa se o número de entradas a_{ij} não nulas é bem inferior a nm ou, equivalentemente, a vasta maioria dos elementos são nulos.

3) Terceira Etapa: Cálculo

Consiste no cálculo da solução, bem como na estimativa da exatidão da mesma.

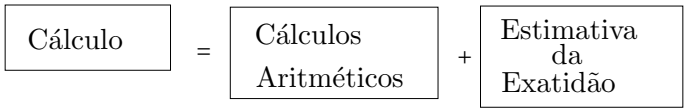


Figura 4.3: Cálculos envolvidos na solução de sistemas de equações lineares.

4.3.1 Primeira Etapa: Descomplexificação

Transformar um sistema complexo $Ax = b$ em um sistema real $\tilde{A}x = \tilde{b}$ equivalente. Ax pode ser escrito como

$$\begin{aligned} Ax = b &\Rightarrow (A' + jA'')(x' + jx'') = (b' + jb'') \\ &\Rightarrow A'x' + jA'x'' + jA''x' - A''x'' = b' + jb'' \end{aligned} \tag{4.2}$$

portanto, o sistema (4.2) pode ser escrito na forma

$$\begin{cases} A'x' - A''x'' = b' \\ A'x'' + A''x' = b'' \end{cases}$$

Exemplo de Aplicação: Análise de Circuitos Elétricos

Para o circuito da Figura 4.4, obtemos através das Leis de Kirchoff:

$$\begin{cases} E - (3-4j)Ri_1 - (2-2j)R(i_1 - i_2) = 0 \\ (2-2j)R(i_2 - i_1) + (1+3j)Ri_2 = 0 \end{cases}$$

Onde E e R são parâmetros, que variam de circuito para circuito. Fazendo:

$$\begin{aligned} i_1 &= \frac{E}{R}x_1 \text{ e} \\ i_2 &= \frac{E}{R}x_2 \end{aligned} \tag{4.3}$$

as equações do circuito podem ser expressas como:

$$\begin{cases} E-(3-4j)R\frac{E}{R}x_1-(2-2j)R(x_1-x_2)\frac{E}{R} &= 0 \\ (2-2j)R(x_2-x_1)\frac{E}{R}+(1+3j)x_2R\frac{E}{R} &= 0 \end{cases} \Rightarrow$$

$$\begin{cases} 1-(3-4j)x_1-(2-2j)(x_1-x_2) &= 0 \\ (2-2j)(x_2-x_1)+(1+3j)x_2 &= 0 \end{cases} \Rightarrow$$

$$\begin{bmatrix} (5-6j) & -(2-2j) \\ -2+2j & 3+j \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Portanto, podemos transformar o sistema acima em um sistema equivalente em variáveis reais, fazendo:

$$A' = \begin{bmatrix} 5 & -2 \\ -2 & 3 \end{bmatrix},\ A'' = \begin{bmatrix} -6 & 2 \\ 2 & 1 \end{bmatrix}$$

$$b' = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ b'' = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ x' = \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix},\ x'' = \begin{bmatrix} x''_1 \\ x''_2 \end{bmatrix},$$

$$\begin{bmatrix} A' & -A'' \\ A'' & A' \end{bmatrix} \begin{bmatrix} x' \\ x'' \end{bmatrix} = \begin{bmatrix} b' \\ b'' \end{bmatrix}$$

Substituindo as matrizes e vetores por seus respectivos valores, obtemos o sistema de equações lineares:

$$\begin{bmatrix} 5 & -2 & 6 & -2 \\ -2 & 3 & -2 & -1 \\ -6 & 2 & 5 & -2 \\ 2 & 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ x''_1 \\ x''_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Exemplo de Aplicação: Regime Permanente de Circuitos Elétricos

Calcular os valores das correntes do circuito dado na Figura 4.5. Inicialmente obtemos a equação diferencial que modela o comportamento dinâmico do sistema:

$$\begin{aligned} V(t) &= RI + L\frac{d}{dt}I + \frac{1}{C}\int_{t=0}^{t=\infty} I \cdot dt \Rightarrow \\ \frac{d}{dt}V &= R\frac{d}{dt}I + L\frac{d^2}{dt^2}I + \frac{1}{C}I \Rightarrow \\ \dot{V} &= L\ddot{I} + R\dot{I} + \frac{1}{C}I \end{aligned}$$

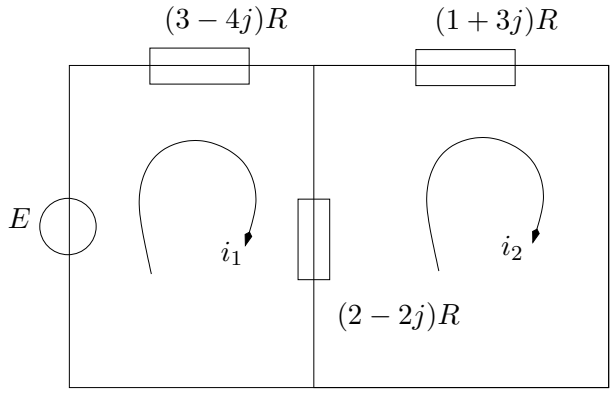


Figura 4.4: Exemplo de circuito elétrico com elementos lineares.

Análise de regime permanente nos dá:

$$\begin{aligned}
\bar{V} &= \bar{V}_r + \bar{V}_c + \bar{V}_l \\
\bar{V} &= R\bar{I} - \frac{j}{\omega C}\bar{I} + j\omega L\bar{I} \\
\bar{V} &= (R - \frac{j}{\omega C} + j\omega L)\bar{I}
\end{aligned}$$

onde $\bar{V}$ é o fasor voltagem e $\bar{I}$ é o fasor corrente. Portanto, podemos expressar o problema de encontrar a corrente e voltagem em regime permanente como a solução de um sistema de equações lineares com coeficientes e variáveis complexas. Primeiro, vamos obter as matrizes e vetores de constantes abaixo:

$$A' = \begin{bmatrix} 5 & -2 \\ -2 & 3 \end{bmatrix}, \quad A'' = \begin{bmatrix} -6 & 2 \\ 2 & 1 \end{bmatrix}, \quad b' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad b'' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Segundo, vamos obter os vetores de variáveis como segue:

$$x' = \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix}, \quad x'' = \begin{bmatrix} x''_1 \\ x''_2 \end{bmatrix}$$

Logo, podemos transformar a busca do ponto de operação do sistema elétrico ao problema de resolver o sistema de equações lineares com coeficientes e variáveis reais:

$$\begin{bmatrix} A' & -A'' \\ A'' & A' \end{bmatrix} \cdot \begin{bmatrix} x' \\ x'' \end{bmatrix} = \begin{bmatrix} b' \\ b'' \end{bmatrix}$$

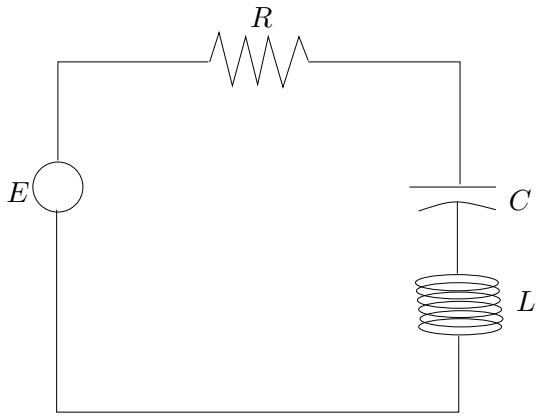


Figura 4.5: Exemplo de circuito elétrico RLC.

4.3.2 Segunda Etapa: Os Algoritmos e Suas Estruturas

Algoritmos Diretos

Estes algoritmos são compostos de três componentes: o calculador, o refinador e o estimador.

Algoritmo Iterativo

São também compostos de três partes: o preparador, o calculador e o estimador.

4.4 Método de Eliminação de Gauss

Método direto mais conhecido e mais usado para resolução de um sistema denso, de pequeno e médio porte, sendo um sistema considerado de pequeno porte se contém até 30 variáveis, médio porte se contém até 50 variáveis, e de grande porte se contém mais de 50 variáveis. O método consiste na aplicação sucessiva de propriedades básicas de álgebra linear.

1) **Combinações lineares:** adição de uma linha com um múltiplo de outra linha, para substituir uma das linhas consideradas.

2) **Troca de linhas**

3) **Multiplicação de uma linha por uma constante**

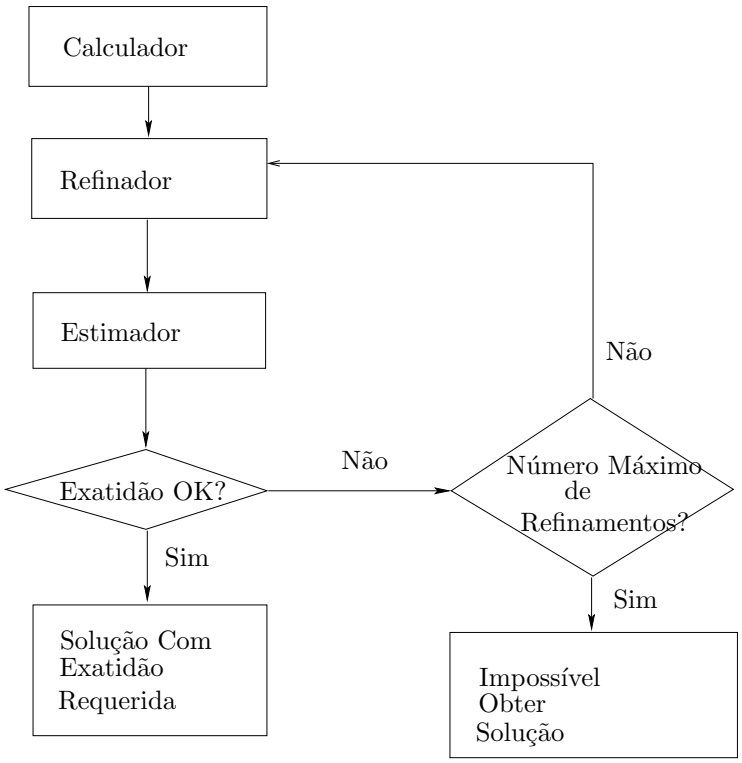


Figura 4.6: Componentes de algoritmos diretos.

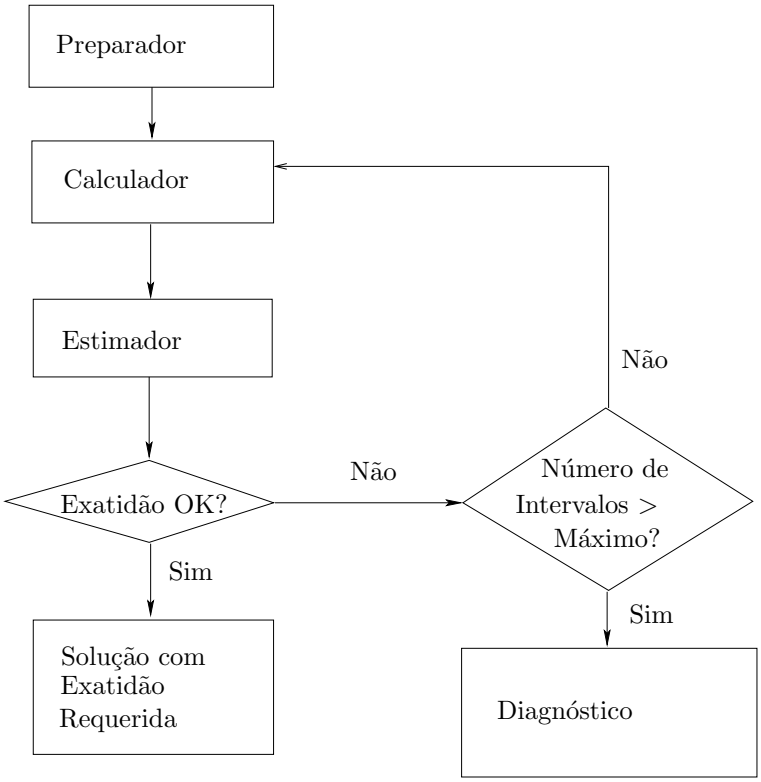


Figura 4.7: Componentes de algoritmos iterativos.

Se a matriz B é obtida a partir de uma matriz A por meio de combinações lineares de linhas, dizemos que A e B são equivalentes. Se A é quadrada então $det(A) = det(B)$.

Algoritmo básico de Gauss apresenta os seguintes passos:

1) **Triangularização:** consiste em transformar a matriz A numa matriz triangular superior, mediante perturbações e combinações lineares de linhas.

2) **Retrossubstituição:** consiste no cálculo dos componentes do vetor x, a partir da solução imediata do último componente de x, e então substituímos regressivamente nas equações anteriores.

4.4.1 Exemplo 1 (Método de Gauss)

Tomemos como exemplo o sistema de equações lineares dado abaixo

$$\begin{cases} 3x_1 + 2x_2 + x_4 = 3 \\ 9x_1 + 8x_2 - 3x_3 + 4x_4 = 6 \\ -6x_1 + 4x_2 - 8x_3 = -16 \\ 3x_1 - 8x_2 + 3x_3 + 4x_4 = 18 \end{cases} \tag{4.4}$$

o qual pode ser escrito na forma matricial como segue

$$A = \begin{bmatrix} 3 & 2 & 0 & 1 \\ 9 & 8 & -3 & 4 \\ -6 & 4 & -8 & 0 \\ 3 & -8 & 3 & 4 \end{bmatrix}, \qquad y = \begin{bmatrix} 3 \\ 6 \\ -16 \\ 18 \end{bmatrix}.$$

A primeira fase consiste da triangularização de A, cujos passos são ilustrados abaixo.

0) Obtendo a matriz aumentada

$$\left[\begin{array}{cccc|c} 3 & 2 & 0 & 1 & 3 \\ 9 & 8 & -3 & 4 & 6 \\ -6 & 4 & -8 & 0 & -16 \\ 3 & -8 & 3 & 4 & 18 \end{array}\right]$$

1) Primeiro passo: zerando os elementos abaixo do elemento a_{11}.

$$\begin{array}{rcl} E_2 - 9/3E_1 & \rightarrow & \\ E_3 + 6/3E_1 & \rightarrow & \\ E_4 - 3/3E_1 & \rightarrow & \end{array} \left[\begin{array}{cccc|c} 3 & 2 & 0 & 1 & 3 \\ 0 & 2 & -3 & 1 & -3 \\ 0 & 8 & -8 & 2 & -10 \\ 0 & -10 & 3 & 3 & 15 \end{array}\right]$$

2) Segundo passo: zerando os elementos abaixo de a_{22}

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} 3 & 2 & 0 & 1 & 3 \\ 0 & 2 & -3 & 1 & -3 \\ 0 & 0 & 4 & -2 & 2 \\ 0 & 0 & -12 & 8 & 0 \end{array}\right] \\ E_3 - 8/2E_2 & \rightarrow & \\ E_4 + 10/2E_2 & \rightarrow & \end{array}$$

3) Terceiro passo: zerando os elementos abaixo de a_{33}

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} 3 & 2 & 0 & 1 & 3 \\ 0 & 2 & -3 & 1 & -3 \\ 0 & 0 & 4 & -2 & 2 \\ 0 & 0 & 0 & 2 & 6 \end{array}\right] \\ E_4 + 12/4E_3 & \rightarrow & \end{array}$$

Obtemos, portanto um sistema equivalente a (4.4) na forma triangular:

$$\begin{cases} 3x_1 + 2x_2 + 0x_3 + x_4 = 3 \\ \quad + 2x_2 - 3x_3 + x_4 = -3 \\ \quad\quad + 4x_3 - 2x_4 = 2 \\ \quad\quad\quad\quad 2x_4 = 6 \end{cases}$$

Por meio de retrossubstituição, podemos encontrar uma solução para o sistema original (4.4). Considerando a última equação temos que

$$2x_4 = 6 \Rightarrow x_4 = 3$$

Substituindo na terceira equação, obtemos:

$$+4x_3 - 2(3) = 2 \Rightarrow x_3 = 2$$

Substituindo na segunda equação, obtemos:

$$+2x_2 - 3 \times 2 + 3 = -3 \Rightarrow x_2 = 0$$

Substituindo na primeira equação, obtemos:

$$3x_1 + 2(0) + 0(2) + 3 = 3 \Rightarrow x_1 = 0$$

Portanto, uma solução para (4.4) é:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 3 \end{pmatrix}$$

Teorema 4.7 *O método de Gauss produz, em precisão infinita, uma solução exata do sistema $Ax = b$ desde que:*

1) A seja não singular, $det(A) \neq 0$

2) As linhas sejam trocadas sempre que necessário, caso $a_{ii} = 0$

4.4.2 Exemplo 2 (Método de Gauss)

Aqui vamos considerar o sistema de equações lineares abaixo:

$$\begin{cases} -1x_1 + 2x_2 + 3x_3 + x_4 = 1 \\ 2x_1 - 4x_2 - 5x_3 - 1x_4 = 0 \\ -3x_1 + 8x_2 + 8x_3 + 1x_4 = 2 \\ 1x_1 + 2x_2 - 6x_3 + 4x_4 = -1 \end{cases} \tag{4.5}$$

Iniciamos a solução pelo método de Gauss com a triangularização do sistema (4.5).

0) Obtendo a matriz aumentada:

$$\left[\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 1 \\ 2 & -4 & -5 & -1 & 0 \\ -3 & 8 & 8 & 1 & 2 \\ 1 & 2 & -6 & 4 & -1 \end{array}\right]$$

1) Primeiro passo: zerando os elementos abaixo de a_{11}

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 4 & -3 & 5 & 0 \end{array}\right] \\ E_2 + 2E_1 & \rightarrow & \\ E_3 - 3E_1 & \rightarrow & \\ E_4 + E_1 & \rightarrow & \end{array}$$

2) Segundo passo: trocando as linhas 2 e 3

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 4 & -3 & 5 & 0 \end{array}\right] \\ E_3 & \rightarrow & \\ E_2 & \rightarrow & \end{array}$$

3) Terceiro passo: zerando os elementos abaixo de a_{22}

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & -1 & 9 & 2 \end{array}\right] \\ E_4 - 2E_2 & \rightarrow & \end{array}$$

4) Quarto passo: zerando os elementos abaixo de a_{33}

$$\begin{array}{rcl} & & \left[\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 10 & 4 \end{array}\right] \\ E_4 + E_3 & \rightarrow & \end{array}$$

Portanto, através de retrossubstituição podemos verificar que a solução de (4.5) é:

$$x_4 = 2/5, \quad x_3 = 8/5, \quad x_2 = 7/10, \quad x_1 = 28/5.$$

4.4.3 Exemplo 3 (Método de Gauss)

Como terceiro exemplo, tomemos os sistema de equações:

$$\begin{cases} 3x_1 + 2x_2 - 1x_3 + 2x_4 = 1 \\ 3x_1 + 4x_2 + 1x_3 + 1x_4 = 3 \\ -6x_1 - 2x_2 + 4x_3 - 3x_4 = 5 \\ -3x_1 - 6x_2 - 3x_3 - 4x_4 = 2 \end{cases} \tag{4.6}$$

Os passos da aplicação do método de Gauss na resolução do sistema (4.6) são descritos no que segue.

0) Obtendo a matriz aumentada:

$$\left[\begin{array}{rrrr|r} 3 & 2 & -1 & 2 & 1 \\ 3 & 4 & 1 & 1 & 3 \\ -6 & -2 & 4 & -3 & 5 \\ -3 & -6 & -3 & -1 & 2 \end{array}\right]$$

1) Primeiro passo: zerando os elementos abaixo de a_{11}

$$\begin{array}{rcl} & & \left[\begin{array}{rrrr|r} 3 & 2 & -1 & 2 & 1 \\ 0 & 2 & 2 & -1 & 2 \\ 0 & 2 & 2 & 1 & 7 \\ 0 & -4 & -4 & 1 & 3 \end{array}\right] \\ E_2 - E_1 & \rightarrow & \\ E_3 + 2E_1 & \rightarrow & \\ E_4 + E_1 & \rightarrow & \end{array}$$

2) Sistema resultante:

$$\begin{array}{rcl} & & \left[\begin{array}{rrrr|r} 3 & 2 & -1 & 2 & 1 \\ 0 & 2 & 2 & -1 & 2 \\ 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & -1 & 7 \end{array}\right] \\ E_3 - E_2 & \rightarrow & \\ E_4 + 2E_2 & \rightarrow & \end{array}$$

A partir das equações obtidas no terceiro passo, verificamos que o sistema não tem solução, pois

$$\begin{cases} 2x_4 = 5 \\ -x_4 = 7 \end{cases}$$

Se o lado direto da terceira equação fosse -14, então as duas últimas equações do sistema reduzido seriam:

$$\begin{cases} 2x_4 &= -14 \\ -x_4 &= 7 \end{cases}$$

Usando este valor $x_4 = -7$ na terceira equação com qualquer valor de x_3, podemos encontrar uma solução para o sistema. Ou seja, o sistema teria um número infinito de soluções.

Teorema 4.8 *Seja $A \in \mathbb{R}^{n\times n}$ uma matriz e $b \in \mathbb{R}^{n\times 1}$ um vetor, então $Ax = b$ pode ser resolvida pelo algoritmo de Gauss com:*

$$\begin{cases} n^2 + \frac{1}{3}(n-1)(n+1)n \in O(n^3) \text{ multiplicações ou divisões} \\ n(n-1) + \frac{1}{6}(n-1)(2n-1)n \in O(n^3) \text{ adições ou subtrações} \end{cases}$$

4.5 Instabilidade Numérica

O algoritmo de Gauss pode apresentar problemas no que diz respeito à exatidão dos resultados. Tais problemas podem ser mais facilmente apresentados por meio de exemplos.

4.5.1 Exemplo de Instabilidade Numérica

Considere o sistema abaixo:

$$\begin{cases} 0.117 \times 10^{-2}x + 0.648y = 0.649 \\ 0.512x - 0.92 \times 10^{-3}y = 0.511 \end{cases} \tag{4.7}$$

Vamos resolver o sistema acima em uma máquina: cujo sistema de ponto flutuante é $F = F(10, 3, -3, 3)$. O arredondamento $\square x$ usado será para o número mais próximo. Primeiro, vamos obter a matriz aumentada de (4.7)

$$\left[\begin{array}{cc|c} 0.117 \times 10^{-2} & 0.648 & 0.649 \\ 0.512 & -0.92 \times 10^{-3} & 0.511 \end{array}\right]$$

Para $M = \frac{-0.512}{0.117\times 10^{-2}} = -437.6068376 \Rightarrow \square(-437.6068376) = -438$, temos que o método que

$$\begin{aligned} M \times 0.648 &= -283.824 \Rightarrow \square(-283.824) = -284 \\ M \times 0.649 &= -284.267 \Rightarrow \square(-284.267) = -284 \end{aligned}$$

Portanto,

$$\begin{aligned} -284 - 0.92 \times 10^{-3} = -284.00092 &\Rightarrow \square(-284.00092) = -284 \\ -284 + 0.511 = -283.489 &\Rightarrow \square(-283.489) = -283 \end{aligned}$$

Após aplicarmos o primeiro passo do método de Gauss, zerando os elementos abaixo de a_{11}, ficamos então com o sistema:

$$\begin{cases} 0.117 \times 10^{-2}x & + & 0.648y & = & 0.649 \\ & - & 0.284 \times 10^{3}y & = & -0.283 \times 10^{3} \end{cases}$$

Portanto, por meio de retrossusbstituição chegamos a solução

$$\begin{cases} x = 0.307 \times 10^{1} = 3.07 \\ y = 0.996 \end{cases}$$

Todavia, obtemos a solução abaixo para o caso de invertermos a ordem das equações:

$$\begin{cases} x = 1 \\ y = 1 \end{cases}$$

A solução correta é a seguinte:

$$\begin{cases} x = 0.999843279\ldots \\ y = 0.99973937\ldots \end{cases}$$

A causa do erro está no pivô pequeno e multiplicador muito grande, o que causa erro de arredondamento. Esta observação leva ao desenvolvimento do método de Gauss com pivoteamento.

4.6 Algoritmo de Gauss com Pivotamento

Este algoritmo nada mais é do que o algoritmo de Gauss com uma troca sistemática de linhas de modo a minimizar os erros de arredondamento.

Estratégia: tome como pivô o elemento de maior valor absoluto. Assim $a_{11} = \max\{|a_{i1}| : i = 1, \ldots, m\}$ será o pivô da primeira iteração; $a_{22} = \max\{|a_{i2}| : i = 2, \ldots, m\}$ será o pivô da segunda iteração e assim por diante. O método a seguir identifica se o sistema tem solução única. As linhas não serão trocadas, mas será criado um vetor que aponta para a linha a ser utilizada: $sub(i) = i$, $i = 1, \ldots, n$. Se as linhas 3 e 7 devem ser trocadas, então: $sub(3) = 7$ e $sub(7) = 3$. Algoritmo da eliminação Gaussiana com pivoteamento segue abaixo.

Gauss_triangularização$(n, a_{ij}, b_i : i = 1, \ldots, n,\ j = 1, \ldots, n)$

1: **for** $i = 1, \ldots, n$ **do**
2: $\quad sub(i) \leftarrow i$

3: **end for**
4: **for** $k = 1, \ldots, n$ **do**
5: $max \leftarrow 0$
6: **for** $i = k, \ldots, n$ **do**
7: $V_{abs} \leftarrow |a_{sub(i),k}|$
8: **if** $max < V_{abs}$ **then**
9: $max \leftarrow V_{abs}$
10: $indx \leftarrow i$
11: **end if**
12: **end for**
13: **if** $max = 0$ **then**
14: Saída "matriz singular"
15: **end if**
16: $j \leftarrow sub(k)$
17: $sub(k) \leftarrow sub(indx)$
18: $sub(indx) \leftarrow j$
19: $pivo \leftarrow a_{sub(k),k}$
20: **for** $i = k + 1, \ldots, n$ **do**
21: $a_{sub(i),k} \leftarrow \frac{a_{sub(i),k}}{pivo}$
22: **for** $j = k + 1, \ldots, n$ **do**
23: $a_{sub(i),j} \leftarrow a_{sub(i),j} - a_{sub(i),k} \times a_{sub(k),j}$
24: **end for**
25: $b_{sub(i)} \leftarrow b_{sub(i)} - a_{sub(i),k} \times b_{sub(k)}$
26: **end for**
27: **end for**

Gauss_retrossubstituição$(n, a_{ij}, b_i : i = 1, \ldots, n,\ j = 1, \ldots, n)$

1: $x_n \leftarrow \frac{b_{sub(n)}}{a_{sub(n),n}}$
2: **for** $k = n - 1, \ldots, 1$ **do**
3: $x_k \leftarrow b_{sub(k)}$
4: **for** $i = k + 1, \ldots, n$ **do**
5: $x_k \leftarrow x_k - a_{sub(k),i} x_i$
6: **end for**
7: $x_k \leftarrow \frac{x_k}{a_{sub(k),k}}$
8: **end for**
9: Saída $x_k : k = 1, \ldots, n$

4.7 Condicionamento de uma Matriz

Seja $Ax = b$ um sistema de equações lineares. Dadas duas aproximações x_1 e x_2, qual delas é melhor? Uma alternativa seria o cálculo dos resíduos:

$$\begin{aligned} r_1 &= b - Ax_1 \\ r_2 &= b - Ax_2 \end{aligned}$$

4.7.1 Exemplo 1

Vamos primeiramente considerar o sistema de equações lineares

$$\begin{cases} 0.24x + 0.636y + 0.12z = 0.84 \\ 0.12x + 0.16y + 0.24z = 0.52 \\ 0.15x + 0.21y + 0.25z = 0.64 \end{cases}$$

Duas aproximações para a solução são:

$$\begin{aligned} x_1 &= [\ 25 \quad -14 \quad -1\ ]^T \\ x_2 &= [\ -3 \quad 4 \quad 0\ ]^T \end{aligned}$$

A partir destes valores, calculamos os resíduos:

$$\begin{aligned} r_1 &= [\ 0.00 \quad 0.00 \quad 0.08\ ]^T \\ r_2 &= [\ 0.12 \quad 0.24 \quad 0.25\ ]^T \end{aligned} \tag{4.8}$$

A solução exata é $x^* = [\ -3 \quad 4 \quad 1\ ]$, embora $\|r_1\| < \|r_2\|$ temos que a solução x_2 é melhor que x_1.

Conclusão: nem sempre a aproximação de menor resíduo é a melhor ou mais exata.

Definição 4.18 *(Problema mal-condicionado) Um problema é dito "mal-condicionado" se pequenas alterações nos dados de entrada ocasionam grandes erros no resultado final.*

4.7.2 Exemplo 2

O sistema de equações lineares abaixo

$$\begin{cases} 0.992x - 0.873y = 0.119 \\ 0.481x - 0.421y = 0.060 \end{cases} \tag{4.9}$$

tem como solução $x = 1$ e $y = 1$. Considere o sistema com uma pequena perturbação ± 0.001, que gera uma aproximação do sistema (4.9):

$$\begin{cases} 0.992x & - & 0.873y & = & 0.120 \\ 0.481x & - & 0.421y & = & 0.060 \end{cases} \tag{4.10}$$

cuja solução é $x = 0.815$ e $y = 0.789$. Um erro de aproximação da ordem 1% na entrada gera um erro de aproximadamente 18% na saída. Mais precisamente:

$$\begin{aligned} E_1 &= \frac{|0.119 - 0.120|}{|0.119|} \\ &= 0.008 \\ &= 1\% \\ E_2 &= \frac{|0.815 - 1|}{|1|} \\ &= 0.185 \\ &= 18\% \end{aligned}$$

4.7.3 Exemplo 3

Considere os sistemas lineares abaixo:

$$\begin{cases} 1x & + & 3y & = & 11 & (a) \\ 1.5x & + & 4.501y & = & 16.503 & (b) \end{cases} \tag{4.11}$$

cuja solução exata é $x = 2$ e $y = 3$. Tomemos agora uma perturbação de (4.11):

$$\begin{cases} 1x & + & 3y & = & 11 & (a) \\ 1.5x & + & 4.501y & = & 16.500 & (c) \end{cases} \tag{4.12}$$

cuja solução é $x = 10.28$ e $y = 0.24$. Os sistemas (4.11) e (4.12) podem ser ilustrados geometricamente, como mostra a Figura 4.8.

4.7.4 Visão Geométrica do Condicionamento

As Figuras 4.9, 4.10, e 4.11 ilustram sistemas sem solução, mal-condicionados e bem-condicionados, respectivamente.

4.7.5 Cálculo do Condicionamento de uma Matriz

Seja o sistema linear $Ax = b$ e uma modificação $Ax' = b'$. Seja x a solução de (A, b) e x' a solução de (A, b'). Qual é a modificação da solução como uma

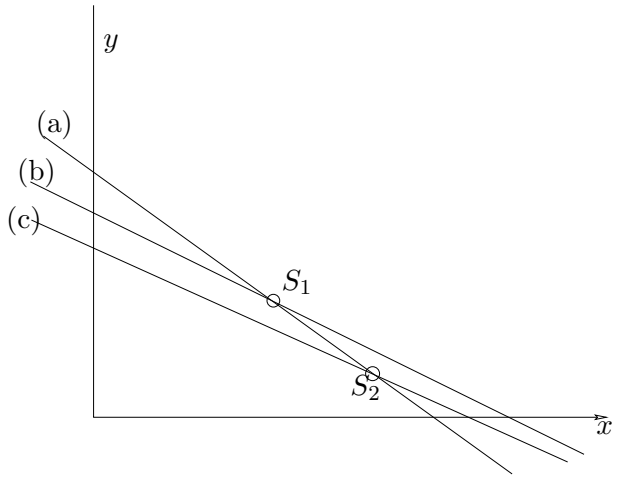


Figura 4.8: Ilustração da solução S_1 do sistema (4.11) e da solução S_2 do sistema (4.12).

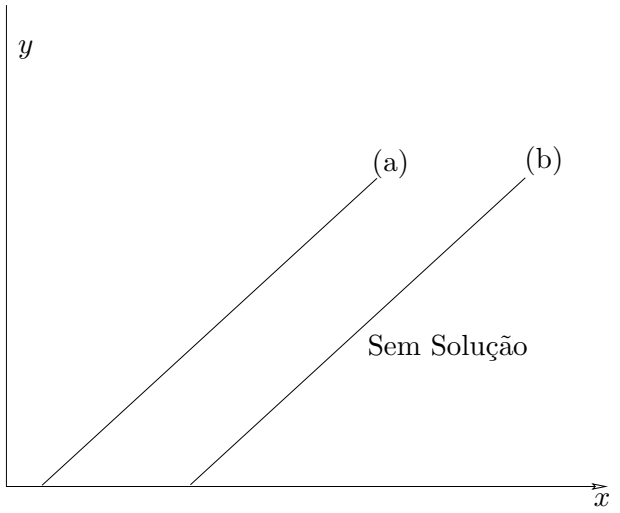


Figura 4.9: Sistema sem solução.

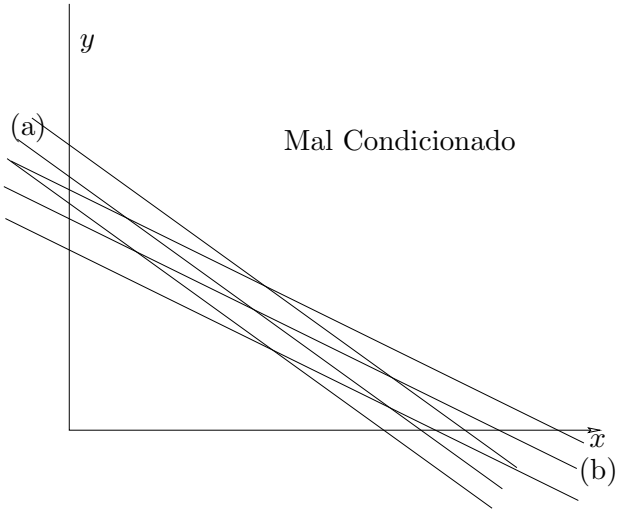


Figura 4.10: Sistema mal-condicionado.

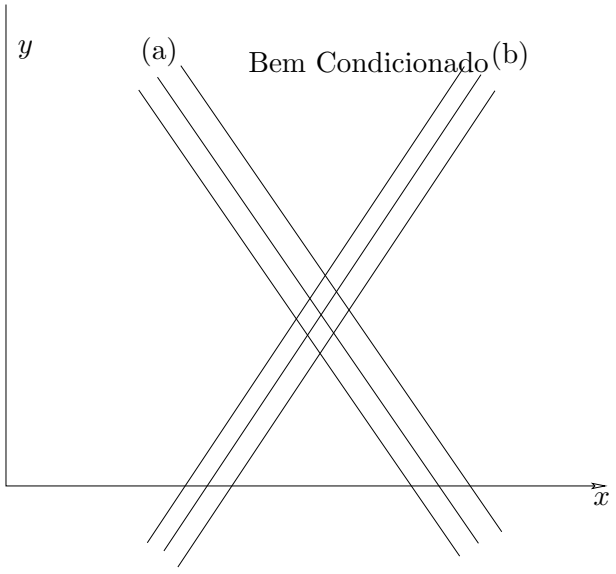


Figura 4.11: Sistema bem-condicionado.

função da modificação em b? Esta resposta pode ser dada de forma algébrica conforme segue:

$$\begin{aligned} b - b' &= Ax - Ax' \\ &= A(x - x') \quad \Rightarrow \\ \\ (x - x') &= A^{-1}(b - b') \end{aligned} \tag{4.13}$$

Tomando a norma de (4.13) temos

$$\|x - x'\| = \|A^{-1}(b - b')\|$$

onde $\|\cdot\|$ é uma norma vetorial. Portanto, pela desigualdade triangular, temos que:

$$\|x - x'\| \leqslant \|A^{-1}\| \cdot \|b - b'\|$$

onde $\|A\|$ é uma norma matricial induzida por $\|\cdot\|$. Assim,

$$\frac{\|x - x'\|}{\|x\|} \leqslant \frac{\|A^{-1}\|}{\|x\|}\|b - b'\|$$

Uma vez que, $b = Ax$, deduzimos o seguinte

$$\begin{aligned} b = Ax &\Rightarrow \|b\| = \|Ax\| \\ &\Rightarrow \|b\| \leqslant \|A\| \cdot \|x\| \\ &\Rightarrow \frac{1}{\|x\|} \leqslant \frac{\|A\|}{\|b\|} \end{aligned} \tag{4.14}$$

Portanto,

$$\begin{array}{ccccc} \text{alteração relativa da solução provocada pela perturbação de } b & \leqslant & \text{fator de amplificação} & \times & \text{valor relativo da perturbação feita no sistema } Ax = b \\ \\ \frac{\|x - x'\|}{\|x\|} & \leqslant & \|A\| \cdot \|A^{-1}\| & \times & \frac{\|b - b'\|}{\|b\|} \end{array} \tag{4.15}$$

Definição 4.19 *Dado um sistema $Ax = b$, seu número de condicionamento é dado por $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ Portanto, quanto maior o valor de $\kappa(A)$, mais sensível o sistema será.*

4.7.6 Exemplo

Tomemos como exemplo o sistema de equações lineares abaixo:

$$\begin{cases} x_1 & + & 10^4 x_2 & = & 10^4 \\ x_1 & + & x_2 & = & 2 \end{cases} \tag{4.16}$$

A partir de (4.16) obtemos a matriz A do sistema e sua inversa A^{-1}:

$$A = \begin{pmatrix} 1 & 10^4 \\ 1 & 1 \end{pmatrix}, \quad A^{-1} = \frac{1}{10^4 - 1} \begin{pmatrix} -1 & 10^4 \\ 1 & -1 \end{pmatrix}$$

Uma vez que $\|A\|_\infty = 10^4 + 1$ e $\|A^{-1}\|_\infty = 1.0002$, verificamos que

$$\begin{aligned} \kappa(A) &= \|A\|_\infty \|A^{-1}\|_\infty \\ &\approx 10^4 \end{aligned}$$

4.7.7 Propriedades da Condicionamento de Matrizes

Algumas propriedades do número de condicionamento de uma matriz A são:

1) $\kappa(A) \geqslant 1$, pois $1 = \|I\| = \|AA^{-1}\| \leqslant \|A\|.\|A^{-1}\| = \kappa(A)$

2) $\kappa(I) = 1$

3) $\forall \alpha \in \mathbb{R}, \kappa(\alpha A) = \kappa(A)$. Demonstração:

$$\begin{aligned} \kappa(\alpha A) &= \|\alpha A\| \cdot \|(\alpha A)^{-1}\| \\ &= |\alpha| \cdot \|A\| \cdot \|\frac{1}{\alpha} A^{-1}\| \\ &= \frac{|\alpha|}{|\alpha|} \|A\| \cdot \|A^{-1}\| \\ &= \kappa(A) \end{aligned}$$

4) Se $D = diag(d_1, d_2, \ldots, d_n)$, então:

$$\kappa(D) = \frac{\max\{|d_j| : j = 1 \ldots, n\}}{\min\{|d_j| : j = 1 \ldots, n\}}$$

Uma definição alternativa do número de condicionamento de uma matriz A é dada por:

$$\kappa(A) = \frac{\max\{\frac{\|Ax\|}{\|x\|} : x \neq 0\}}{\min\{\frac{\|Ax\|}{\|x\|} : x \neq 0\}}$$

Observação: Dígitos Significativos

Aplicando o logaritmo nos dois lados da desigualdade (4.15), obtemos:

$$\begin{aligned} \log \frac{\|x - x'\|}{\|x\|} &\leqslant \log \kappa(A) \frac{\|b - b'\|}{\|b\|} \\ &= \log \kappa(A) + \log \frac{\|b - b'\|}{\|b\|} \end{aligned} \tag{4.17}$$

Por meio de aproximação, a desigualdade acima nos leva a:

$$-DIGSE(x') \leqslant \log \kappa(A) - DIGSE(b')$$

e, portanto,

$$DIGSE(x') \geqslant DIGSE(b') - \log \kappa(A)$$

Se $\kappa(A) = 10^j$, então j dígitos significativos poderão ser perdidos.

4.8 Refinamento da Solução para o Método de Gauss

Conforme visto, o cálculo do condicionamento é difícil e oneroso do ponto de vista computacional. O método de Gauss, mesmo com pivoteamento, não produz em geral nenhuma estimativa sobre a exatidão da resposta. Com a técnica dos refinamentos podemos obter uma medida da exatidão da resposta, bem como avaliar se o sistema dado é bem ou mal condicionado.

4.8.1 Descrição do Método

Os passos do método de refinamento são dados na seqüência.

Passo 1: obter uma primeira aproximação x^1 da solução exata de $Ax = b$ (via Gauss com pivoteamento).

Passo 2: Refinar a solução obtida a partir de x^k, gerando uma aproximação x^{k+1} e obtendo mediante condições de convergência $x = \lim_{k \to \infty} x^k$.

4.8.2 Geração de Aproximações

A partir de x^1, queremos determinar z^1, tal que:

$$x^1 + z^1 = x, \text{ onde } Ax = b$$

Portanto, z_1 é a diferença entre a solução exata x e a solução aproximada x_1.

Primeiro Refinamento: Determinar z^1

O refinamento é dado por:

$$x^1 + z^1 = x \Rightarrow z^1 = x - x^1$$

e multiplicando por A os dois lados da igualdade, temos

$$\begin{aligned} Az^1 &= A(x - x^1) \\ &= Ax - Ax^1 \\ &= b - Ax^1 \\ &= r^1 \end{aligned}$$

Note que z^1 pode ser calculado resolvendo os sistema de equações:

$$Az^1 = r^1$$

Note que z^1 pode ser calculado com facilidade. Só precisamos aplicar as operações que levaram à triangularização de A ao vetor de resíduos r^1. Obtemos então $x^2 = x^1 + \square z^1$.

Segundo Refinamento: Determinar z^2

Determinar o segundo refinamento, z^2, consiste em encontrar:

$$\begin{aligned} z^2 = x - x^2 &\Rightarrow Az^2 = A(x - x^2) \\ &\Rightarrow Az^2 = b - Ax^2 = r^2 \end{aligned}$$

Assim, calculamos z^2 e obtemos:

$$x^3 = x^2 + \square z^2$$

Obtemos, portanto, uma sequência $x^1, x^2, \ldots, x^k$, para a qual veremos uma condição de convergência.

Teorema 4.9 *Seja $Ax = b$ um sistema de equações lineares e x^k a sequência obtida por refinamentos. Se:*

(a) $\kappa(A) < \frac{1}{16\mu(n^3+3n^2)}$

(b) os resíduos r^k são calculados com precisão dupla.

Então x^k converge para a solução exata.

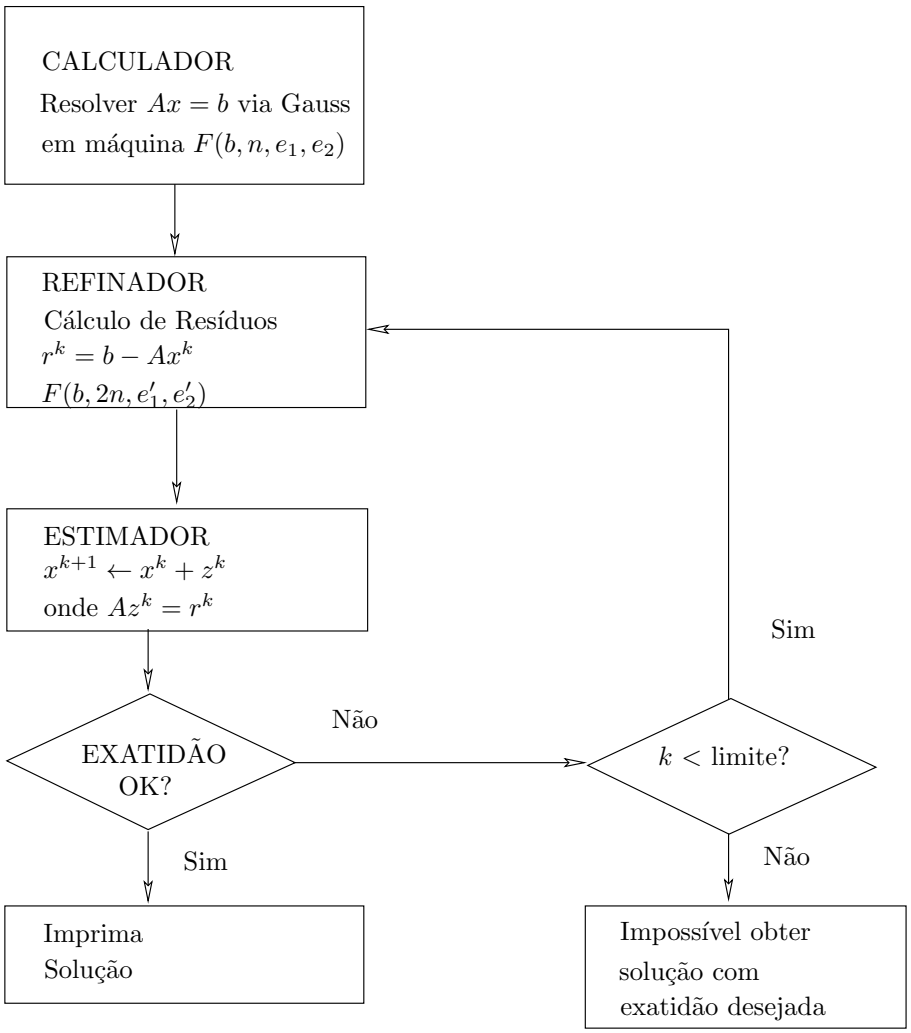


Figura 4.12: Algoritmo de refinamento.

4.8.3 Algoritmo

Para aplicar o método dos refinamentos, precisamos resolver o sistema $Az^k = r^k$. O algoritmo é descrito na Figura 4.12.

4.8.4 Exemplo

Para fins de exemplo, considere o sistema de equações lineares abaixo:

$$\begin{cases} 2.4579x_1 + 1.6235x_2 + 4.6231x_3 = 0.064700 \\ 1.4725x_1 + 0.9589x_2 - 1.3253x_3 = 1.0473 \\ 2.6951x_1 + 2.8965x_2 - 1.4794x_3 = -0.6789 \end{cases} \tag{4.18}$$

em uma máquina de ponto flutuante dada por $F = F(10, 5, -98, 100)$. Os passos do algoritmo são descritos no que segue.

1) Triangularização: Inicialmente, vamos triangularizar a matriz aumentada. A matriz aumentada $[A|b]$ é dada por:

$$\left[\begin{array}{ccc|c} 2.4579 & 1.6235 & 4.6231 & 0.0647 \\ 1.4725 & 0.9589 & -1.3253 & 1.0473 \\ 2.6951 & 2.8965 & -1.4794 & -0.6789 \end{array}\right]$$

Trocando as linhas 1 e 3, obtemos

$$\begin{array}{rc} E_3 & \to \\ & \\ E_1 & \to \end{array} \left[\begin{array}{ccc|c} 2.6951 & 2.8965 & -1.4794 & -0.6789 \\ 1.4725 & 0.9589 & -1.3253 & 1.0473 \\ 2.4579 & 1.6235 & 4.6231 & 0.0647 \end{array}\right]$$

Zerando os elementos abaixo de a_{11}, obtemos:

$$\begin{array}{rc} E_2 - 1.4725/2.6951E_1 & \to \\ E_3 - 2.4579/2.6951E_1 & \to \end{array} \left[\begin{array}{ccc|c} 2.6951 & 2.8965 & -1.4794 & -0.6789 \\ 0 & -0.6236 & -0.51702 & 1.4182 \\ 0 & -1.0374 & 5.9822 & 0.68839 \end{array}\right]$$

$$\begin{array}{rc} E_3 & \to \\ E_2 & \to \end{array} \left[\begin{array}{ccc|c} 2.6951 & 2.8965 & -1.4794 & -0.6789 \\ 0 & -1.0374 & 5.9822 & 0.68839 \\ 0 & -0.6236 & -0.51702 & 1.4182 \end{array}\right]$$

$$E_3 \to E_3 - \tfrac{-0.6236}{-1.0374}E_2 \to \left[\begin{array}{ccc|c} 2.6951 & 2.8965 & -1.4794 & -0.6789 \\ 0 & -1.0374 & 5.9822 & 0.68839 \\ 0 & 0 & -4.1132 & 1.0044 \end{array}\right]$$

Logo, temos a matriz A em sua forma triangular superior.

2) Retrossubstituição: Fazendo a retrossubstituição, calculamos a solução aproximada para (4.18):

$$\begin{cases} x_3 &= -0.24419 \\ x_2 &= -2.0717 \\ x_1 &= 1.8406 \end{cases}$$

Seja $x^1 = [1.8406 - 2.0717 - 0.24419]^T$ a primeira solução aproximada.

2) Refinamento: em $F(10, 10, -98, 100)$

Utilizando aproximação x^1 calculamos

$$Ax^1 = \begin{pmatrix} 0.064821801 \\ 1.047355377 \\ -0.678823304 \end{pmatrix}$$

o que nos permite calcular o resíduo

$$r^1 = b - Ax^1 = \begin{bmatrix} 0.0648 \\ 1.0473 \\ -0.6789 \end{bmatrix} - \begin{bmatrix} 0.064821801 \\ 1.047355377 \\ -0.678823304 \end{bmatrix} = \begin{bmatrix} -0.000121801 \\ -0.000055377 \\ -0.000076696 \end{bmatrix}$$

Resolvemos $Az^1 = r^1$ usando as trocas e os multiplicadores já calculados, obtendo z^1 e arredondando os valores de z^1 para precisão simples. Temos então:

$$z^1 = \begin{pmatrix} z_1^1 \\ z_2^1 \\ z_3^1 \end{pmatrix} = \begin{pmatrix} -0.000004228 \\ 0.000025110 \\ -0.000057765 \end{pmatrix}$$

Assim, podemos obter a segunda aproximação $x^2 = x^1 + z^1$ dada por:

$$x^2 = \begin{bmatrix} 1.8405 \\ -2.0717 \\ -0.24419 \end{bmatrix}$$

4.8.5 Análise do Condicionamento Através do Refinamento

Para análise do comportamento dos sistemas lineares sem o uso do condicionamento de A, podemos detectar o mal condicionamento sem calcular $\kappa(A)$ diretamente ou aproximadamente. Se os resíduos $r^1, r^2, \ldots, r^k$, são

pequenos, mas as correções $z^1, z^2, \ldots, z^k$, são grandes, então o sistema é *mal-condicionado*. Para sistemas *bem-condicionados*, os refinamentos não devem ser feitos mais do que duas vezes.

A avaliação empírica de sistemas lineares mal-condicionados pode ser feita através de um gráfico, conforme ilustram as Figuras 4.13 (sistema bem-condicionado), 4.14 (sistema mais ou menos mal-condicionado), e 4.15 (sistema mal-condicionado).

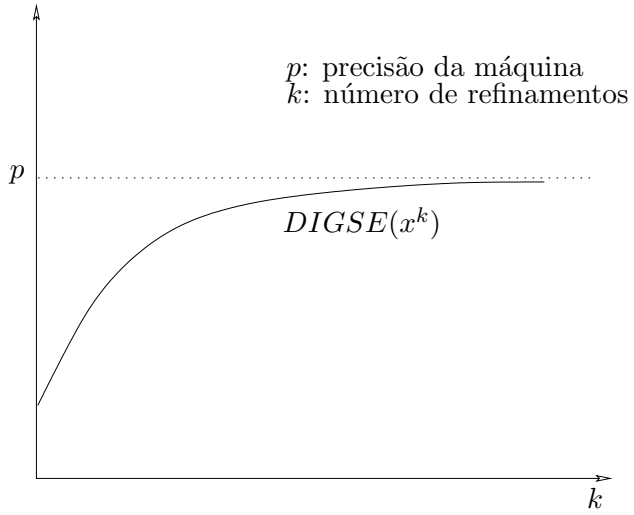


Figura 4.13: Avaliação empírica de sistema bem-condicionado através do método de refinamento.

4.9 Equacionamento Matricial: Eliminação Gaussiana de Forma Compacta

Vamos construir matrizes que expressam a eliminação Gaussiana em termos matriciais. O equacionamento será ilustrado através de um exemplo. Seja $Ax = b$ um sistema de equações lineares dado por:

$$A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{bmatrix} \text{ e } b = \begin{bmatrix} 7 \\ 4 \\ 6 \end{bmatrix}$$

Fazendo:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix},$$

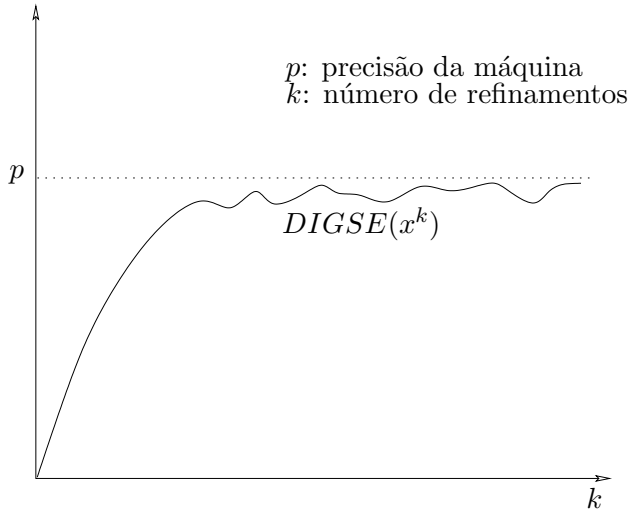


Figura 4.14: Avaliação empírica de sistema mais ou menos mal-condicionado através do método de refinamento.

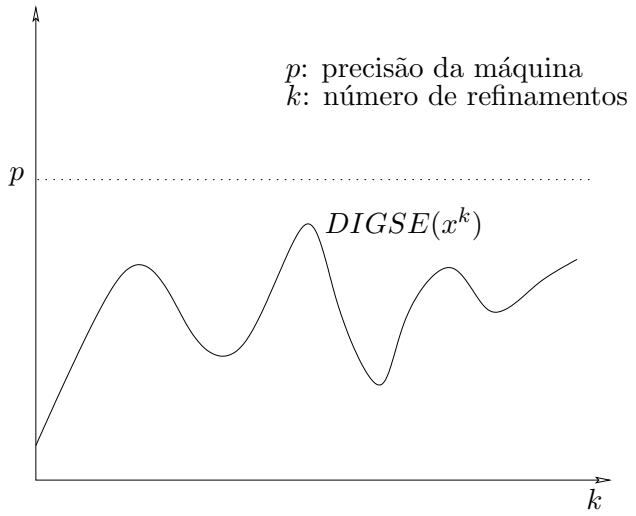


Figura 4.15: Avaliação empírica de sistema mal-condicionado através do método de refinamento.

onde $m_{21} = 0.3$ e $m_{31} = -0.5$, obteremos:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0.3 & 1 & 0 \\ -0.5 & 0 & 1 \end{bmatrix}$$

Portanto,

$$M_1 Ax = M_1 b \;\Leftrightarrow\; \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} x = \begin{bmatrix} 7 \\ 6.1 \\ 2.5 \end{bmatrix}$$

Em outras palvras, a pré-multiplicação de $Ax = b$ com a matriz M_1 resultou no zeramento dos elementos abaixo de a_{11}. Repetindo os passos acima para o elemento da coluna 2, abaixo de a_{22}, executaremos o segundo passo da triangularização de A. Seja então:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix},$$

onde $m_{32} = 25$, o que nos leva a:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 25 & 1 \end{bmatrix}$$

Pré-multiplicando $M_1 Ax = M_1 b$ com a matriz M_2, obteremos

$$M_2(M_1 A)x = M_2(M_1)b \;\Leftrightarrow\; M_2(M_1 A) = \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{bmatrix} x = \begin{bmatrix} 7 \\ 6.1 \\ 155 \end{bmatrix}$$

Portanto, o método de Gauss pode ser escrito como:

$$M_2 M_1 Ax = M_2 M_1 b$$

onde $M_2 M_1 A = U$ é uma matriz triangular superior. Em geral, para um sistema $A \in \mathbb{R}^{n \times n}$, temos

$$(M_{n-1} \ldots M_2 M_1)Ax = (M_{n-1} \ldots M_2 M_1)b$$

O que acontece quando temos que trocar linhas, durante o pivoteamento? Seja P_{ij} a matriz obtida ao trocarmos as linhas i e j da matriz identidade I. Então, $P_{ij}A$ troca as linhas i e j da matriz A.

Exemplo

Sejam as matrizes

$$A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{bmatrix}, \text{ e } P_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

então $P_{23}A$ é precisamente

$$P_{23}A = \begin{bmatrix} 10 & -7 & 0 \\ 5 & -1 & 5 \\ -3 & 2 & 6 \end{bmatrix}$$

Equacionamento do Matricial do Método de Gauss

Portanto, o método de Gauss pode ser expresso como:

$$\begin{aligned} U &= M_{n-1}P_{n-1}M_{n-2}P_{n-2}\ldots M_2P_2M_1P_1A \\ &\Rightarrow \\ A &= P_1^{-1}M_1^{-1}P_2^{-1}M_2^{-1}\ldots P_{n-1}^{-1}M_{n-1}^{-1}U \\ &= LU \\ &\Rightarrow \\ A &= P_1^TM_1^{-1}P_2^TM_2^{-1}\ldots P_{n-1}^TM_{n-1}^{-1}U \\ &= LU \end{aligned}$$

Uma vez que $P_j^{-1} = P_j^T$ sempre que P_j for uma matriz de permutação de linhas. Note que M_i^{-1} pode ser facilmente obtida a partir de M_i: a matriz inversa M_i^{-1} é idêntica à matriz M_i, exceto que o sinal dos elementos abaixo da diagonal principal têm sinais opostos aos elementos correspondentes em M_i.

Exemplo de $P_j^{-1} = P_j^T$

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

Exercício

i. Calcule as matrizes inversas M_1^{-1} e M_2^{-1} do exemplo. Verifique a propriedade que relaciona M_i^{-1} a M_i. Obtenha a representação $A = LU$.

4.10 Decomposição LU

Dado um sistema de equações lineares, com A não-singular,

$$A \cdot x = b \tag{4.19}$$

a decomposição LU busca encontrar três matrizes $n \times n$ de maneira que

$$P \cdot A = L \cdot U \tag{4.20}$$

onde:

- P é uma matriz de pivoteamento (também indicada na literatura como matriz de permutação);
- L é uma matriz triangular inferior (Lower); e
- U é uma matriz triangular superior (Upper).

Este método busca transformar o problema em dois problemas fáceis de serem resolvidos, que é a resolução de sistemas lineares com matrizes triangulares. De posse das matrizes PLU, o processo de resolução é:

$$\begin{aligned} A \cdot x &= b \\ P^{-1} \cdot L \cdot U \cdot x &= b \end{aligned} \tag{4.21}$$

Então definimos $y = Ux$ e pré-multiplicamos a equação (4.21) por P, obtendo:

$$\begin{aligned} P \cdot P^{-1} \cdot L \cdot y &= P \cdot b \\ L \cdot y &= P \cdot b \end{aligned} \tag{4.22}$$

O sistema na equação (4.22) pode ser resolvido por substituição direta, o que nos leva a encontrar y. Assim, possuímos elementos para resolver:

$$U \cdot x = y \tag{4.23}$$

por retrosubstituição.

4.10.1 Substituição Direta

Admita o sistema de equações lineares:

$$L \cdot y = b$$

onde L é uma matriz diagonal inferior. Usaremos um exemplo com $L \in \mathbb{R}^{3\times3}$ para exemplificar o processo, onde:

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Podemos observar que

$$y_1 = \frac{b_1}{l_{11}}$$

e obtemos seu valor de imediato. No passo seguinte temos que

$$\begin{aligned} l_{21}y_1 + l_{22}y_2 &= b_2 \\ y_2 &= \frac{1}{l_{22}}(b_2 - l_{21}y_1) \end{aligned}$$

mas y_1 já é conhecido, então pode-se obter o valor de y_2. Desta maneira, y_3 e eventuais y_n podem ser calculados sucessivamente.

4.10.2 Retrosubstituição

Admita agora o sistema de equações lineares:

$$U \cdot x = y$$

onde U é uma matriz diagonal superior. Utilizamos novamente um exemplo com $U \in \mathbb{R}^{3\times3}$ para apresentar o processo, onde:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Da mesma maneira que na seção anterior, mas iniciando a partir da última linha de U, temos que:

$$\begin{aligned} x_3 &= \frac{y_3}{u_{33}} \\ x_2 &= \frac{1}{u_{22}}(y_2 - u_{23}x_3) \\ &\vdots \end{aligned}$$

4.10.3 Obtendo LU sem Permutações

Primeiro consideraremos o problema da obtenção da fatoração sem a matriz de permutações. Este caso especial é obtido fazendo $P = I$ (identidade) no caso geral. Admita a seguinte divisão das matrizes $A = L \cdot U$:

$$\begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}$$

onde A_{21} e $L_{21} \in \mathbb{R}^{(n-1)\times 1}$ são matrizes colunas; A_{12} e $U_{12} \in \mathbb{R}^{1\times(n-1)}$ são matrizes linhas; e A_{22}, L_{22}, $U_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$ agrupam o restante das matrizes originais. Do produto obtemos:

$$\begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} u_{11} & U_{12} \\ u_{11}L_{21} & L_{21}U_{12} + L_{22}U_{22} \end{bmatrix}$$

Assim podemos obter a primeira linha de U e a primeira coluna de L, fazendo:

$$\begin{aligned} u_{11} &= a_{11} \\ U_{12} &= A_{12} \\ L_{21} &= \frac{1}{u_{11}} A_{21} \end{aligned} \tag{4.24}$$

e ainda obtemos:

$$L_{22}U_{22} = A_{22} - L_{21}U_{12} = A_{22} - \frac{1}{a_{11}} A_{12}A_{21} \tag{4.25}$$

que nos permite calcular L_{22} e U_{22} fatorando $A_{22} - L_{21}U_{12}$. Assim, chegamos ao seguinte algoritmo recursivo:

1) calcular a primeira linha de U: $u_{11} = a_{11}$ e $U_{12} = A_{12}$;

2) calcular a primeira coluna de L: $L_{12} = (\frac{1}{a_{11}})A_{21}$;

3) calcular a decomposição das sub-matrizes L_{22} e U_{22}: $L_{22}U_{22} = A_{22} - L_{21}U_{12}$.

Este algoritmo fica da seguinte maneira em pseudo-código:

LU-Solve(A)
1: $n \leftarrow rows[A]$
2: **for** $k \leftarrow 1$ to n **do**
3: $\quad u_{kk} \leftarrow a_{kk}$
4: $\quad$ **for** $i \leftarrow k+1$ to n **do**

```
5:       l_ik ← a_ik/u_kk
6:       u_ki ← a_ki
7:     end for
8:     for i ← k + 1 to n do
9:       for j ← k + 1 to n do
10:        a_ij ← a_ij − l_ik u_kj
11:      end for
12:    end for
13: end for
14: Retorna L e U
```

O algoritmo assume que a matriz U já possui zeros abaixo de sua diagonal, e que a matriz L possui zeros acima de sua diagonal e 1's nela.

Exemplo

Calcular a fatoração LU de

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 6 & 13 & 5 \\ 2 & 19 & 10 \end{bmatrix}$$

It $k = 1$:

$$\left[\begin{array}{c|cc} 2 & 3 & 1 \\ \hline 6 & 4 & 2 \\ 2 & 16 & 9 \end{array}\right] \Longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & & \\ 0 & 0 & \end{bmatrix}$$

It $k = 2$:

$$\left[\begin{array}{cc|c} 2 & 3 & 1 \\ 6 & 4 & 2 \\ \hline 2 & 16 & 1 \end{array}\right] \Longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & \end{bmatrix}$$

It $k = 3$:

$$\begin{bmatrix} 2 & 3 & 1 \\ 6 & 4 & 2 \\ 2 & 16 & 1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

Discussão do Algoritmo

Podemos observar que o algoritmo modifica a matriz A e que os valores contidos nas linhas e colunas k só são utilizados em sua iteração, assim é possível que as matrizes L e U sejam montadas na própria matriz A.

4.10.4 Problemas do Não-Pivoteamento

Um exemplo simples mostra que nem todas as matrizes não-singulares podem ser fatoradas por $A = LU$:

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

Se aplicarmos o algoritmo, teremos $u_{11} = 0$. Para calcularmos $L_{21} = \frac{1}{u_{11}} A_{21}$, teremos uma divisão por zero. Assim, elementos nulos na posição do "pivô" devem ser evitados utilizando, por exemplo, uma troca de linhas.

De fato, a troca de linhas é bastante comum, procurando que o pivô assuma o elemento de maior norma da coluna restante. Esta heurística se mostra bastante eficiente para manter a estabilidade numérica do algoritmo, salvo em casos raros.

4.10.5 Representando as Matrizes LU em Uma Matriz

Pela sua definição, a matriz U apresenta elementos não nulos acima de sua diagonal e nela. A matriz L os possui abaixo de sua diagonal e nela. Uma conseqüência desta implementação é que os elementos da diagonal principal de L sempre serão 1. Assim, podemos representar as duas matrizes em uma só, pois vários elementos já são conhecidos de antemão:

$$A' = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & u_{22} & u_{23} & & u_{2n} \\ l_{31} & l_{32} & u_{33} & & u_{3n} \\ \vdots & & & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & u_{nn} \end{bmatrix}$$

4.10.6 O Vetor de Pivoteamento

No algoritmo que será apresentado a seguir, é utilizado um vetor π ao invés da matriz para representar a permutação entre linhas. Estas notações são equivalentes. No vetor, o elemento $\pi[i]$ indica qual é a linha da matriz I que deve estar no local de i.

Exemplo

O vetor $\pi = (2, 4, 1, 3)$ é equivalente à matriz

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

4.10.7 Decomposição LU com Pivoteamento

A demostração formal do algoritmo com pivoteamento é um pouco mais complicada que do primeiro, e não será coberta neste material. Foi demonstrado que esta decomposição existe sempre se a matriz A for não-singular. Nesta implementação a decomposição será executada na própria matriz A, obtendo a representação apresentada na seção 4.10.5. As permutações serão representadas pelo vetor π.

O algoritmo possui quatro etapas distintas: as três primeiras linhas são a inicialização das variáveis e vetores; da linha 5 à 11 a coluna é varrida a partir de k, na busca do elemento de maior norma; nas três linhas seguintes é executada a operação de troca de linhas; por fim, nas linhas 15 à 18 são calculados os novos fatores da matriz. O algoritmo em pseudo-código é:

LUP-Solve(A)

1: $n \leftarrow rows[A]$
2: **for** $i \leftarrow 1$ até n **do**
3: $\quad \pi[i] \leftarrow i$
4: **end for**
5: **for** $k \leftarrow 1$ até n **do**
6: $\quad p \leftarrow 0$
7: $\quad$ **for** $i \leftarrow k$ até n **do**
8: $\quad\quad$ **if** $|a_{ik}| > p$ **then**
9: $\quad\quad\quad p \leftarrow |a_{ik}|$
10: $\quad\quad\quad k' \leftarrow i$
11: $\quad\quad$ **end if**
12: $\quad\quad$ **if** $p = 0$ **then**
13: $\quad\quad\quad$ Saída "matriz singular"
14: $\quad\quad\quad$ Pare
15: $\quad\quad$ **end if**
16: $\quad\quad$ Troque $\pi[k] \leftrightarrow \pi[k']$
17: $\quad\quad$ **for** $i \leftarrow 1$ até n **do**
18: $\quad\quad\quad$ Troque $a_{ki} \leftrightarrow a_{k'i}$

```
      end for
      for i ← k + 1 até n do
        a_ik ← a_ik/a_kk
        for j ← k + 1 até n do
          a_ij ← a_ij − a_ik a_kj
        end for
      end for
    end for
end for
Retorne A e π
```

Exemplo

Calcular a fatoração LUP da matriz

$$A = \begin{bmatrix} 2 & 0 & 2 & 0,6 \\ 3 & 3 & 4 & -2 \\ 5 & 5 & 4 & 2 \\ -1 & -2 & 3,4 & -1 \end{bmatrix}$$

Iteração $k = 1$:

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \begin{matrix} 2 & 0 & 2 & 0,6 \\ 3 & 3 & 4 & -2 \\ (5) & 5 & 4 & 2 \\ -1 & -2 & 3,4 & -1 \end{matrix} \qquad \begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \begin{matrix} (5) & 5 & 4 & 2 \\ 3 & 3 & 4 & -2 \\ 2 & 0 & 2 & 0,6 \\ -1 & -2 & 3,4 & -1 \end{matrix}$$

$$\begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \begin{matrix} (5) & 5 & 4 & 2 \\ 0,6 & 0 & 1,6 & -3,2 \\ 0,4 & -2 & 0,4 & -0,2 \\ -0,2 & -1 & 4,2 & -0,6 \end{matrix}$$

Iteração $k = 2$:

$$\begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,6 & 0 & 1,6 & -3,2 \\ 0,4 & (-2) & 0,4 & -0,2 \\ -0,2 & -1 & 4,2 & -0,6 \end{matrix} \qquad \begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,4 & (-2) & 0,4 & -0,2 \\ 0,6 & 0 & 1,6 & -3,2 \\ -0,2 & -1 & 4,2 & -0,6 \end{matrix}$$

$$\begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,4 & (-2) & 0,4 & -0,2 \\ 0,6 & 0 & 1,6 & -3,2 \\ -0,2 & 0,5 & 4 & -0,5 \end{matrix}$$

Iteração $k = 3$:

$$\begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,4 & -2 & 0,4 & -0,2 \\ 0,6 & 0 & 1,6 & -3,2 \\ -0,2 & 0,5 & (4) & -0,5 \end{matrix} \qquad \begin{bmatrix} 3 \\ 1 \\ 4 \\ 2 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,4 & -2 & 0,4 & -0,2 \\ -0,2 & 0,5 & (4) & -0,5 \\ 0,6 & 0 & 1,6 & -3,2 \end{matrix}$$

$$\begin{bmatrix} 3 \\ 1 \\ 4 \\ 2 \end{bmatrix} \begin{matrix} 5 & 5 & 4 & 2 \\ 0,4 & -2 & 0,4 & -0,2 \\ -0,2 & 0,5 & (4) & -0,5 \\ 0,6 & 0 & 0,4 & -3 \end{matrix}$$

Assim,

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & 2 & 0,6 \\ 3 & 3 & 4 & -2 \\ 5 & 5 & 4 & 2 \\ -1 & -2 & 3,4 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0,4 & 1 & 0 & 0 \\ -0,2 & 0,5 & 1 & 0 \\ 0,6 & 0 & 0,4 & 1 \end{bmatrix} \begin{bmatrix} 5 & 5 & 4 & 2 \\ 0 & -2 & 0,4 & -0,2 \\ 0 & 0 & 4 & -0,5 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

4.10.8 Método de Crout

Aqui apresentamos um algoritmo alternativo para encontrar a decomposição LU sem pivoteamento. Podemos obter LU resolvendo um sistema de equações não lineares conforme segue:

$$\underbrace{\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}}_{U} = \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}}_{A}$$

Isto significa que

$$a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{in}u_{nj}$$

Alguns exemplos de equações:

$$\begin{aligned} a_{13} &= l_{11}u_{13} + l_{12}u_{23} + l_{13}u_{33} + l_{14}u_{43} \\ &= l_{11}u_{13} \end{aligned}$$

$$\begin{aligned} a_{32} &= l_{31}u_{12} + l_{32}u_{22} + l_{33}u_{32} + l_{34}u_{42} \\ &= l_{31}u_{12} + l_{32}u_{22} \end{aligned}$$

$$\begin{aligned} a_{22} &= l_{21}u_{12} + l_{22}u_{22} + l_{23}u_{32} + l_{24}u_{42} \\ &= l_{21}u_{12} + l_{22}u_{12} \end{aligned}$$

Há três casos:

$$1)\ i < j: \quad a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ii}u_{ij} \tag{4.26}$$
$$2)\ i = j: \quad a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ii}u_{ij} \tag{4.27}$$
$$3)\ i > j: \quad a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{ij}u_{jj} \tag{4.28}$$

O sistema de equações (4.26), (4.27), e (4.28) possui n^2 equações e $n^2 + n$ variáveis (a diagonal é representada duas vezes). Uma vez que o número de variáveis é maior que o número de incógnitas, podemos especificar n das incógnitas arbitrariamente. Então fazemos,

$$l_{ii} = 1,\ \forall i = 1, \ldots, n \tag{4.29}$$

O algoritmo de Crout, que será apresentado a seguir, resolve o sistema de $N^2 + N$ equações (4.26), (4.27), (4.28), e (4.29) através de uma simples reordenação das equações.

Crout(A)
1: Defina $l_{ii} = 1$, $i = 1, \ldots, n$
2: **for** $j = 1, \ldots, n$ **do**
3: **for** $i = 1, \ldots, j$ **do**
4: Use (4.26), (4.27), e (4.29) para resolver u_{ij}: $u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}$
5: **end for**
6: **for** $i = j+1, j+2, \ldots, n$ **do**
7: Use (4.28) para resolver l_{ij}: $l_{ij} = \frac{1}{u_{jj}}(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj})$
8: **end for**
9: **end for**

Exemplo

Aqui vamos aplicar o Algoritmo de Crout para encontrar a decomposição LU da matriz

$$A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

Passos executados pelo algoritmo:

1) $u_{11} = a_{11} - \sum_{k=1}^{1-1} l_{ik}u_{kj} = a_{11} = 10$

2) $l_{21} = \frac{1}{u_{11}}(a_{21} - \sum_{k=1}^{1-1} l_{ik}u_{kj}) = -0.3$

3) $l_{31} = \frac{1}{u_{11}}(a_{31} - \sum_{k=1}^{1-1} l_{ik}u_{kj}) = 0.5$

4) $u_{12} = a_{12} - \sum_{k=1}^{1-1} l_{ik}u_{kj} = a_{12} = -7$

5) $u_{22} = a_{22} - \sum_{k=1}^{2-1} l_{ik}u_{kj} = a_{22} - l_{21}u_{12} = 2 - (0.3)(-7) = 2 - 2.1 = -0.1$

6) $l_{32} = \frac{1}{u_{22}}(a_{32} - \sum_{k=1}^{2-1} l_{3k}u_{k2}) = -25$

7) $u_{13} = a_{13} - \sum_{k=1}^{1-1} l_{ik}u_{kj} = a_{13} = 0$

8) $u_{23} = a_{23} - \sum_{k=1}^{2-1} l_{2k}u_{k3} = a_{23} - l_{31}u_{13} = 6 - (0.3)(0) = 6$

9) $u_{33} = a_{33} - \sum_{k=1}^{3-1} l_{3k}u_{k3} = a_{33} - l_{31}u_{13} - l_{32}u_{23} = 5 - (0.5)(0) - (-25)6 = 155$

Portanto,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -0.3 & 1 & 0 \\ 0.5 & -25 & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{bmatrix}$$

o que nos permite verificar que

$$LU = \begin{bmatrix} 1 & 0 & 0 \\ -0.3 & 1 & 0 \\ 0.5 & -25 & 1 \end{bmatrix} \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{bmatrix} = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

4.11 Decomposição de Cholesky

Aqui consideramos uma variante da fatoração LU aplicada a matrizes simétricas e positiva definidas. Seja A uma matriz simétrica e positiva definida, ou seja $A = A^T$ e $A \succ 0$, então existem uma matriz triangular inferior L tal que $A = LL^T$. Para se calcula a matriz L, inicialmente colocamos a matriz A na forma abaixo:

$$A = \begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}$$

onde a_{11} é um escalar, A_{21} é uma matriz linha, A_{21} é um vetor coluna e A_{22} é uma submatriz de dimensão apropriada. Note que o escalar $a_{11} > 0$ visto que $A \succ 0$. Então o fator Cholsky L pode ser colocado em uma forma apropriada

$$L = \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}$$

onde l_{11} é um escalar, L_{21} é um vetor coluna e L_{22} é uma matriz de mesma dimensão de A_{22}. Logo,

$$\begin{aligned} A = \begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} &= \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} l_{11}^2 & l_{11} L_{21}^T \\ l_{11} L_{21} & (L_{21} L_{21}^T + L_{22} L_{22}^T) \end{bmatrix} = LL^T \end{aligned}$$

Os passos para obtenção do fator Cholesky são:

i. Calcule a primeira coluna de L:

$$\begin{aligned} l_{11}^2 = a_{11} &\implies l_{11} = \sqrt{a_{11}} \\ l_{11} L_{21} = A_{21} &\implies L_{21} = \frac{1}{l_{11}} A_{21} \end{aligned}$$

ii. Recursivamente compute a fatoração de Cholesky da submatriz

$$L_{21} L_{21}^T + L_{22} L_{22}^T = A_{22} \implies L_{22} L_{22}^T = A_{22} - L_{21} L_{21}^T$$

Sendo o termo $L_{21} L_{21}^T$ conhecido, enquanto o $L_{22} L_{22}^T$ desconhecido, podemos expressar a fatoração Cholesky da submatriz como segue

$$\begin{aligned} L_{22} L_{22}^T &= A_{22} - L_{21} L_{21}^T \\ &= A_{22} - \frac{1}{l_{11}} A_{21} A_{21}^T \\ &= A_{22} - \frac{1}{\sqrt{a_{11}}} A_{21} A_{21}^T \end{aligned}$$

Exemplo

$$A = \begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & l_{22} & 0 \\ -1 & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} 5 & 3 & -1 \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

$$\begin{bmatrix} 18 & 0 \\ 0 & 11 \end{bmatrix} = \begin{bmatrix} l_{22} & 0 \\ l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{22} & l_{32} \\ 0 & l_{33} \end{bmatrix} + \begin{bmatrix} 3 \\ -1 \end{bmatrix} [3 \quad -1]$$

Portanto,

$$\begin{bmatrix} 9 & 3 \\ 3 & 10 \end{bmatrix} = \begin{bmatrix} l_{22} & 0 \\ l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{22} & l_{32} \\ 0 & l_{33} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & l_{33} \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & l_{33} \end{bmatrix}$$

$$10 = 1 + l_{33}^2 \rightarrow l_{33} = 3$$

o que nos leva ao fator Cholesky

$$L = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 3 & 0 \\ -1 & 1 & 3 \end{bmatrix}.$$

4.12 Método de Gauss-Jordan

O método de Gauss-Jordan transforma a matriz A do sistema $Ax = b$ em uma matriz diagonal.

1a Etapa: transformar $Ax = b$ em $Ux = \bar{b}$, onde U é uma matriz triangular superior.

2a Etapa: transformar o sistema $Ux = \bar{b}$ em um sistema $Dx = \hat{b}$ onde D é uma matriz diagonal.

Observação: a primeira e segunda etapas podem ser realizadas simultaneamente. A k-ésima equação será usada para zerar os coeficientes das equações $1, 2, \ldots, k-1, k+1, \ldots, n$.

Exemplo

Abaixo seguem os passos envolvidos na primeira etapa do método de Gauss-Jordan. Vamos considerar a matriz aumentada $[A|b]$ de um sistema

de equações lineares $Ax = b$.

$$\left[\begin{array}{rrr|c} 1 & -3 & 4 & 1 \\ -2 & 8 & -6 & 2 \\ 3 & -5 & 15 & 5 \end{array}\right] \rightarrow \left[\begin{array}{rrr|c} 1 & -3 & 4 & 1 \\ 0 & 2 & 2 & 4 \\ 0 & 4 & 3 & 2 \end{array}\right]$$
$$\rightarrow \left[\begin{array}{rrr|r} 1 & 0 & 7 & 7 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & -1 & -6 \end{array}\right]$$
$$\rightarrow \left[\begin{array}{rrr|r} 1 & 0 & 0 & -35 \\ 0 & 2 & 0 & -8 \\ 0 & 0 & -1 & -6 \end{array}\right]$$

Com base no sistema equivalente obtido na primeira etapa, podemos deduzir que a solução para $Ax = b$ é dada por:

$$x_1 = -35,\ x_2 = -4,\ x_3 = 6$$

Podemos utilizar o método de Gauss-Jordan para obter a inversa da matriz.

$$A = \left[\begin{array}{rrr} 1 & -3 & 4 \\ -2 & 8 & -16 \\ 3 & -5 & 15 \end{array}\right]$$

Para isto, tomamos a matriz aumentada $[A|I]$ e aplicamos os mesmos passos da primeira etapa, conforme indicado abaixo:

$$\left[\begin{array}{rrr|ccc} 1 & -3 & 4 & 1 & 0 & 0 \\ -2 & 8 & -16 & 0 & 1 & 0 \\ 3 & -5 & 15 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{rrr|ccc} 1 & -3 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 2 & 1 & 0 \\ 0 & 4 & 3 & -3 & 0 & 1 \end{array}\right]$$
$$\rightarrow \left[\begin{array}{rrr|ccc} 1 & 0 & 7 & 4 & \frac{3}{2} & 0 \\ 0 & 1 & 1 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & -1 & -7 & -2 & 1 \end{array}\right]$$
$$\rightarrow \left[\begin{array}{rrr|ccc} 1 & 0 & 0 & -45 & -\frac{25}{2} & 7 \\ 0 & 1 & 0 & -6 & -\frac{3}{2} & 1 \\ 0 & 0 & 1 & 7 & 2 & -1 \end{array}\right]$$

Logo,

$$A^{-1} = \left[\begin{array}{rrr} -45 & -\frac{25}{2} & 7 \\ -6 & -\frac{3}{2} & 1 \\ 7 & 2 & -1 \end{array}\right]$$

Os passos executados acima podem ser descritos de outra forma

$$\begin{aligned} A^{-1}[A|I] &\rightarrow [A^{-1}A|A^{-1}I] \\ &\rightarrow [I|A^{-1}] \end{aligned}$$

4.13 Método de Gauss-Jordan

O método de Gauss-Jordan transforma a matriz A do sistema $Ax = b$ em uma matriz diagonal.

1a Etapa: transformar $Ax = b$ em $Ux = \bar{b}$, onde U é uma matriz triangular superior.

2a Etapa: transformar o sistema $Ux = \bar{b}$ em um sistema $Dx = \hat{b}$ onde D é uma matriz diagonal.

Observação: a primeira e segunda etapas podem ser realizadas simultaneamente. A k-ésima equação será usada para zerar os coeficientes das equações $1, 2, \ldots, k-1, k+1, \ldots, n$.

Exemplo

Abaixo seguem os passos envolvidos na primeira etapa do método de Gauss-Jordan. Vamos considerar a matriz aumentada $[A|b]$ de um sistema de equações lineares $Ax = b$.

$$\left[\begin{array}{rrr|r} 1 & -3 & 4 & 1 \\ -2 & 8 & -6 & 2 \\ 3 & -5 & 15 & 5 \end{array}\right] \Rightarrow \left[\begin{array}{rrr|r} 1 & -3 & 4 & 1 \\ 0 & 2 & 2 & 4 \\ 0 & 4 & 3 & 2 \end{array}\right]$$

$$\Rightarrow \left[\begin{array}{rrr|r} 1 & 0 & 7 & 7 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & -1 & -6 \end{array}\right]$$

$$\Rightarrow \left[\begin{array}{rrr|r} 1 & 0 & 0 & -35 \\ 0 & 2 & 0 & -8 \\ 0 & 0 & -1 & -6 \end{array}\right]$$

Com base no sistema equivalente obtido na primeira etapa, podemos deduzir que a solução para $Ax = b$ é dada por:

$$x_1 = -35,\ x_2 = -4,\ x_3 = 6$$

Podemos utilizar o método de Gauss-Jordan para obter a inversa da matriz.

$$A = \left[\begin{array}{rrr} 1 & -3 & 4 \\ -2 & 8 & -16 \\ 3 & -5 & 15 \end{array}\right]$$

Para isto, tomamos a matriz aumentada $[A|I]$ e aplicamos os mesmos passos da primeira etapa, conforme indicado abaixo:

$$\left[\begin{array}{rrr|ccc} 1 & -3 & 4 & 1 & 0 & 0 \\ -2 & 8 & -16 & 0 & 1 & 0 \\ 3 & -5 & 15 & 0 & 0 & 1 \end{array}\right] \Rightarrow \left[\begin{array}{rrr|rrr} 1 & -3 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 2 & 1 & 0 \\ 0 & 4 & 3 & -3 & 0 & 1 \end{array}\right]$$

$$\Rightarrow \left[\begin{array}{rrr|rrr} 1 & 0 & 7 & 4 & \frac{3}{2} & 0 \\ 0 & 1 & 1 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & -1 & -7 & -2 & 1 \end{array}\right]$$

$$\Rightarrow \left[\begin{array}{rrr|rrr} 1 & 0 & 0 & -45 & -\frac{25}{2} & 7 \\ 0 & 1 & 0 & -6 & -\frac{3}{2} & 1 \\ 0 & 0 & 1 & 7 & 2 & -1 \end{array}\right]$$

Logo,

$$A^{-1} = \left[\begin{array}{rrr} -45 & -\frac{25}{2} & 7 \\ -6 & -\frac{3}{2} & 1 \\ 7 & 2 & -1 \end{array}\right]$$

Os passos executados acima podem ser descritos de outra forma

$$\begin{aligned} A^{-1}[A|I] &\Rightarrow [A^{-1}A|A^{-1}I] \\ &\Rightarrow [I|A^{-1}] \end{aligned}$$

4.14 *Singular Value Decomposition* (SVD)

SVD é uma técnica poderosa para tratar de conjuntos de equações ou matrizes que são singulares ou quase singulares. Em muitos casos SVD não apenas faz um diagnóstico do problema, mas também encontra uma solução, no sentido de que SVD encontra uma solução numérica útil. Seja $A \in \mathbb{R}^{m\times n}$ uma matriz com m linhas e n colunas, $m \geqslant n$, então A pode ser escrita como:

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

onde:

- $U \in \mathbb{R}^{m\times n}$, cujas colunas são ortogonais,
- $\Sigma \in \mathbb{R}^{n\times n}$ é a diagonal cujas entradas são números positivos ou zeros (valores singulares)
- $V \in \mathbb{R}^{n\times n}$, que é ortogonal.

Em outras palavras

$$A = U \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \sigma_n \end{bmatrix} V^T = U \cdot diag(\sigma_1, \dots, \sigma_n) \cdot V^T$$

tal que $U^T U = V^T V = diag(1, 1, \dots, 1)$

Para $A \in \mathbb{R}^{n \times n}$, temos

$$A = U\Sigma V^T,\ U \in \mathbb{R}^{n \times n},\ V \in \mathbb{R}^{n \times n} \text{ e } \Sigma \in \mathbb{R}^{n \times n}$$

SVD também pode ser aplicado quando $m < n$. Neste caso os valores singulares σ_j, $j = m+1, \dots, n$, são todos nulos e as colunas correspondentes de U são nulas.

A decomposição SVD pode ser executada sempre, não importando quão singular a matriz seja e a decomposição é quase-única. A decomposição SVD é única a menos de permutações das colunas de U, dos elementos de Σ, e colunas de V (ou linhas de V^T)

4.14.1 SVD de uma Matriz Quadrada

Seja $A \in \mathbb{R}^{n \times n}$ uma matriz quadrada $A = U\Sigma V^T$ a decomposição SVD correspondente. Se A é não-singular, podemos calcular sua inversa

$$\begin{aligned} A^{-1} &= (U\Sigma V^T)^{-1} \\ &= (\Sigma V^T)^{-1} U^{-1} \\ &= (V^T)^{-1} \Sigma^{-1} U^{-1} \\ &= V\Sigma^{-1} U^T \\ &= V diag(1/\sigma_1, 1/\sigma_2, \dots, 1/\sigma_n) U^T \end{aligned}$$

Note que

$$\begin{aligned} AA^{-1} &= (U\Sigma V^T)(U\Sigma V^T)^{-1} \\ &= (U\Sigma V^T) V\Sigma^{-1} U^T \\ &= U\Sigma (V^T V) \Sigma^{-1} U^T \\ &= U(\Sigma\Sigma^{-1}) U^T \\ &= UU^T \\ &= I \end{aligned}$$

A inversa de A não pode ser calculada quando σ_j é próximo de zero, para algum j, o que indica que a matriz é quase-singular. A condição de A, $\kappa(A)$, pode ser calculada a partir da diagonal de Σ conforme segue

$$\kappa(A) = \frac{\max\{\sigma_j : j = 1, \ldots, n\}}{\min\{\sigma_j : j = 1, \ldots, n\}}$$

Dizemos que matriz A é singular se $\kappa(A) = \infty$ e mal-condicionada de $\kappa(A)$ é um número grande.

Observação: O algoritmo para cálculo de decomposição SVD pode ser encontrado em pacotes numéricos, tais como Matlab, Scilab e Mathematica. O algoritmo de decomposição SVD é intrincado e por esta razão não será apresentado neste texto.

4.14.2 Determinando o Espaço Gerado e o Espaço Nulo

Podemos determinar o espaço gerado ($range(A)$) e o espaço nulo ($null(A)$) de uma matriz singular A através da decomposição SVD. Seja o sistema $Ax = b$, onde $A \in \mathbb{R}^{n\times n}$ é uma matriz quadrada. Lembramos que $null(A) = \{x \in \mathbb{R}^n : Ax = 0\}$, $range(A) = \{y \in \mathbb{R}^n : y = Ax \text{ para algum } x\}$, e $dim(range(A)) + dim(null(A)) = n$.

O que SVD tem a ver com $null(A)$ e $range(A)$? SVD explicitamente produz bases para $null(A)$ e $range(A)$.

- As colunas de U cujos w_j correspondentes são não nulos, formam uma base ortonormal de $range(A)$.

- As colunas de V cujos w_j correspondentes são nulos formam uma base ortonormal para o espaço nulo.

4.14.3 Exemplo 1

Consideremos o sistema de equações lineares $Ax = b$ dado por

$$\begin{bmatrix} 1 & -3 & 4 \\ -2 & 8 & -6 \\ 3 & -5 & 15 \\ -1 & 5 & -2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

A decomposição SVD nos dá $A = U\Sigma V^T$, onde

$$
\begin{aligned}
U &= \begin{bmatrix} 0.2594 & -0.0876 & -0.7692 \\ -0.4840 & 0.6269 & 0.1987 \\ 0.8050 & 0.5555 & 0.2083 \\ -0.2245 & 0.5393 & -0.5705 \end{bmatrix} \\
\Sigma &= \begin{bmatrix} 19.5468 & 0 & 0 \\ 0 & 6.0765 & 0 \\ 0 & 0 & 0.0292 \end{bmatrix} \\
V &= \begin{bmatrix} 0.1978 & -0.0352 & 0.9796 \\ -0.5012 & 0.8552 & 0.1320 \\ 0.8424 & 0.5171 & -0.1515 \end{bmatrix}
\end{aligned} \tag{4.30}
$$

Com base na decomposição SVD acima, podemos verificar que o número de condição da matriz A é $\kappa(A) = 670.2104$.

4.14.4 Exemplo 2

Aqui ilustramos como a decomposição SVD pode ser empregada para se encontrar o espaço gerado e o espaço nulo de uma matriz. Considere a matriz:

$$
B = \begin{bmatrix} 1 & -3 & 4 & -2 \\ -2 & 8 & -6 & 6 \\ 3 & -5 & 15 & -2 \\ 1 & 2 & -2 & 3 \end{bmatrix}
$$

A decomposição SVD de $B = U\Sigma V^T$ é dada pelas matrizes:

$$
\begin{aligned}
U &= \begin{bmatrix} 0.2707 & 0.1120 & -0.0693 & 0.9536 \\ -0.5352 & -0.7314 & 0.3316 & 0.2619 \\ 0.7834 & -0.6011 & 0.0554 & -0.1477 \\ -0.1627 & -0.3019 & -0.9392 & 0.0134 \end{bmatrix} \\
\Sigma &= \begin{bmatrix} 20.0342 & 0 & 0 & 0 \\ 0 & 6.7846 & 0 & 0 \\ 0 & 0 & 1.8975 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
V &= \begin{bmatrix} 0.1761 & -0.0782 & -0.7934 & -0.5774 \\ -0.4660 & -0.5580 & 0.3716 & -0.5774 \\ 0.8172 & -0.5271 & 0.2333 & 0.0000 \\ -0.2899 & -0.6362 & -0.4218 & 0.5774 \end{bmatrix}
\end{aligned}
$$

Note que B é singular, $\kappa(B) = \infty$. Com base na decomposição SVD acima, podemos calcular uma base para o espaço nulo de B.

$$Null(B) = range(\begin{bmatrix} -0.5774 \\ -0.5774 \\ 0.0000 \\ 0.5774 \end{bmatrix})$$

e também uma base para o espaço gerado por B

$$Range(B) = Range(\begin{bmatrix} 0.2707 & 0.1120 & -0.0693 \\ -0.5352 & -0.7314 & 0.3316 \\ 0.7834 & -0.6011 & 0.0554 \\ -0.1627 & -0.3019 & -0.9392 \end{bmatrix})$$

4.14.5 Considerações Finais sobre Decomposição SVD

Considere o sistema $Ax = b$, onde $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^{n \times 1}$, e $x \in \mathbb{R}^{n \times 1}$. Suponha que A é singular, ou seja, $det(A) = 0$, e $b \in range(A)$. Neste caso, existem infinitas soluções x que satisfazem $Ax = b$. Em particular, seja x tal que $Ax = b$, então $A(x + y) = b$, $\forall y \in null(A)$. Considere o problema abaixo:

$$\begin{array}{rll} P: & \text{Minimize} & \|x\|^2 \\ & x \in \mathbb{R}^n & \\ & \text{Sujeito a :} & \\ & & Ax = b \end{array}$$

Em palavras, o problema P se refere a encontrar a solução x para $Ax = b$ que tenha a menor norma. Seja $\Sigma^{-1} = diag(\frac{1}{\sigma_j})$, fazendo $\frac{1}{\sigma_j} = 0$ sempre que o valor singular w_j for nulo. Então $x^* = V diag(\frac{1}{\sigma_j}) U^T b$ é a solução ótima de P.

4.14.6 Visão Geométrica da Decomposição SVD

Transformações lineares apresentam várias aplicações. Considere uma matriz $A \in \mathbb{R}^{n \times n}$ e a transformação linear $f(x) = Ax$. Note que dois vetores de mesma direção são mapeados em vetores de mesma direção pela função $f(x)$. Como exemplo, considere os vetores x e $x' = \alpha x$ com $\alpha \in \mathbb{R} - \{0\}$. Então $f(x) = Ax$ e $f(x') = Ax' = A(\alpha x) = \alpha Ax$ o que nos permite entender a transformação a partir da visão geométrica da transformação de vetores unitários. Uma transformação linear (não-singular) A transforma esferas em

elipsóides. A Fig. 4.16 ilustra a transformação induzida pela matriz A dada por:

$$A = \begin{bmatrix} 1.47 & 0.98 \\ 0.98 & 1.47 \end{bmatrix} \tag{4.31}$$

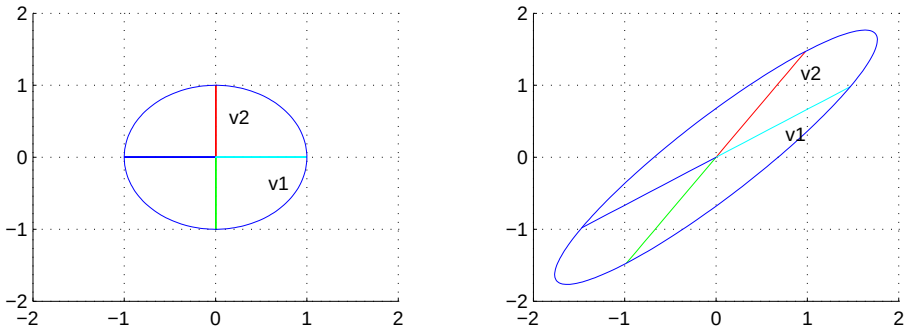


Figura 4.16: Transformação de esferas em elipsóides.

Uma maneira de entender o que A faz é encontrar quais vetores são mapeados para os eixos principais do elipsóide. Se tivermos sorte, $A = U\Lambda U^T$ com U ortogonal, o que ocorre quando A é simétrica, então os autovetores de A definem os eixos do elipsóide. A decomposição por autovetores de A nos diz quais vetores ortogonais têm comprimento redimensionado (ver Fig. 4.17):

$$A = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}^T \tag{4.32}$$

$$Av_i = \lambda_i v_i \tag{4.33}$$

Note que $f(v_i) = Av_i = \lambda_i v_i$, ou seja, um autovetor v_i matém a direção sofrendo apenas um ajuste no comprimento (e sentido, caso $\lambda_i < 0$) conforme o autovalor λ_i.

Como exemplo, a decomposição por autovetores da matriz A dada em (4.31) pode ser expressa como segue:

$$A = U\Lambda U^T = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} 0.49 & 0 \\ 0 & 2.45 \end{bmatrix} \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \tag{4.34}$$

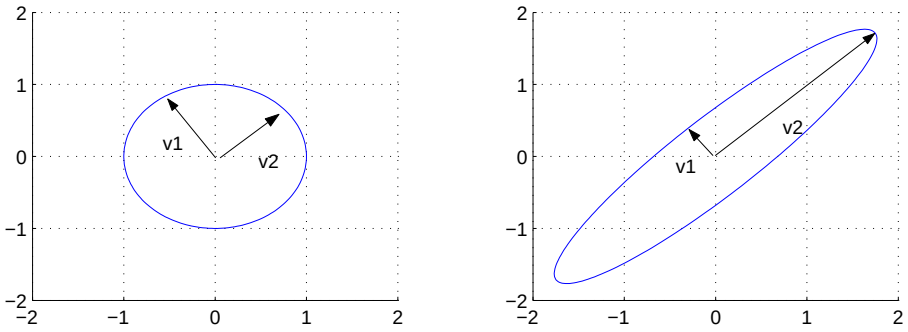


Figura 4.17: Ilustração da transformação induzida por uma matriz A simétrica.

Esta transformação é ilustrada na Fig. 4.18. Note que o vetor $v_1 = [-0.7071\ 0.7071]$ também matém direção e sentido, tendo seu comprimento reduzido pela metade conforme fator $\lambda_1 = 0.49$. Por outor lado, o vetor $v_2 = [0.7071\ 0.7071]$ mantém a direção e sentido, mas o comprimento é expandido pelo fator $\lambda_2 = 2.45$.

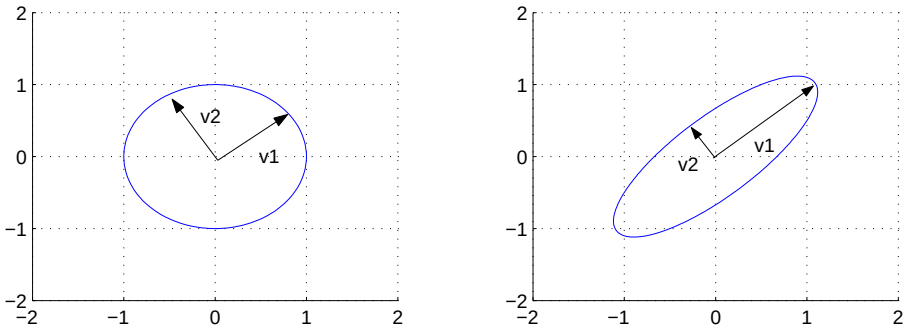


Figura 4.18: Ilustração da transformação induzida por uma matriz A simétrica.

Em geral uma matriz A também induz rotações, não apenas alterações nas escalas como indica a Fig. 4.19. A transformação dada por A pode ser

melhor entendida através da decomposição SVD:

$$A = U\Sigma V^T \tag{4.35}$$

$$= \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}^T \tag{4.36}$$

onde as matrizes U e V são ortonormais e $\sigma_i \geqslant 0$ são os valores singulares. Logo, temos que $AV = U\Sigma V^T V = U\Sigma$. Para um vetor v_i a transformação dada por A leva a um vetor $Av_i = \sigma_i u_i$:

$$Av_i = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}^T v_i \tag{4.37}$$

$$= \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ \sigma_i \\ \vdots \\ 0 \end{bmatrix} = \sigma_i u_i \tag{4.38}$$

Para qualquer matriz quadrada $A \in \mathbb{R}^{n\times n}$ existem matrizes ortogonais $U, V \in \mathbb{R}^{n\times n}$ e uma matriz diagonal Σ, tal que os elementos σ_i da diagonal de Σ são não negativos com $A = U\Sigma V^T$. Os elementos da diagonal de Σ $(\sigma_1, \dots, \sigma_n)$ são ditos valores singulares e tipicamente se assume que eles estão ordenados com $\sigma_1 \geqslant \sigma_2 \geqslant \sigma_n$. As colunas de U $(u_1, \dots, u_n)$ são chamadas de vetores singulares à esquerda e definem os eixos do elipsóide (ver Fig. 4.19). As colunas de V $(v_1, \dots, v_n)$ são chamadas de vetores singulares à direita e definem os vetores que sujeitos à transformação levam aos eixos do elipsóide, $Av_i = \sigma_i u_i$.

4.14.7 Análise dos Componentes Principais

Considere um experimento com um número n de pessoas onde se mede a altura de cada indivíduo em centímetros e polegadas. Neste experimento é avaliado o impacto de suplementos nutricionais na altura. É necessário manter ambas as medidas de altura? Provavelmente não, pois a altura é

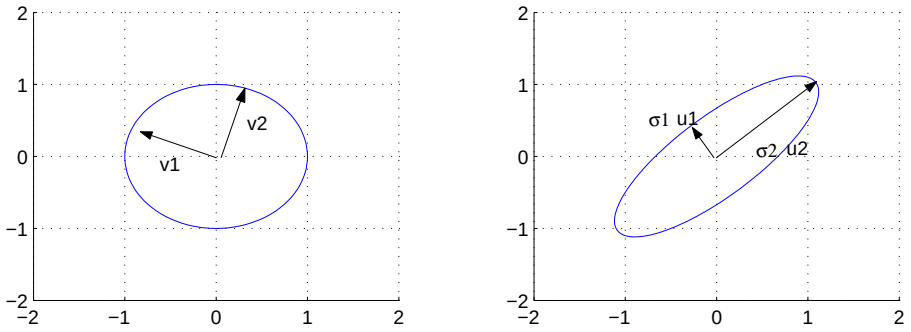


Figura 4.19: Ilustração da transformação dada por uma matriz A que gera rotações e mudanças em escala.

uma característica de uma pessoa independentemente de como ela é medida, seja em centímetros, seja em polegadas.

Em uma situação mais complexa, um questionário poderia ser elaborado para avaliar a satisfação de pessoas com suas vidas. Neste questionário há questões sobre os atividades de lazer *(item 1)* e quão freqüentemente dedicam tempo para recreação e lazer *(item 2)*. As chances desses dois itens estarem correlacionados são muito altas. Existindo alto grau de correlação (medida estatística de dependência entre duas variáveis aleatórias), concluímos que os itens são redudantes.

Podemos sumarizar a dependência entre duas variáveis através de um gráfico, contendo em uma dimensão uma das variáveis e na outra dimensão, a outra variável. Uma reta pode então ser ajustada para representar um relacionamento linear entre as variáveis, como ilustra a Fig. 4.20. Se pudéssemos definir uma variável que aproximasse a reta em tal gráfico, então estaríamos capturando a essência dos dois itens. O questionário aplicado aos sujeitos do experimento acima poderia tomar como base este novo fator, representado pela linha, que captura a essência dos *itens* 1 e 2. De certa maneira, reduzimos as duas variáveis a um único fator. Observe que o novo fator é na verdade uma combinação linear das duas variáveis.

O exemplo de combinar duas variáveis correlacionadas em um único fator ilustra a idéia básica por trás da análise de componentes principais. Estendendo o caso de duas variáveis para o cenário de múltiplas variáveis, então o cômputo dos fatores principais se torna mais complexo, mas o princípio de modelar duas ou mais variáveis por um único fator permanece. Em linhas gerais, a extração dos componentes principais consiste em uma rotação que maximiza variância (*variance maximization rotation*) aplicada ao espaço nativo das variáveis. Para o exemplo dado acima, visto na Fig. 4.20, pode-

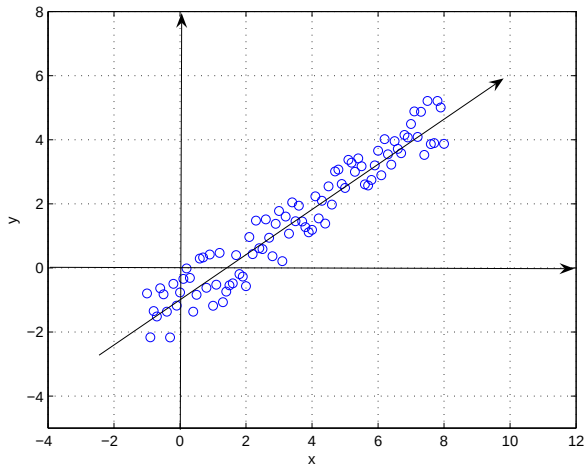


Figura 4.20: Regressão linear entre duas variáveis.

mos imaginar uma translação dos pontos de forma que o centro de massa se torne a origem, seguida de uma rotação do eixo x em torno da origem que aproxime a reta de regressão. Este tipo de rotação é dita *maximizadora de variância* pois o critério da rotação é maximizar a variância (*variabilidade*) do novo fator, enquanto minimiza a variância em torno da nova variável. Este procedimento é ilustrado nas Figs. 4.21 e 4.22.

Na presença de mais de duas variáveis, podemos entendê-las como definindo um "espaço", da mesma forma que duas variáveis definem um plano. Quando temos três variáveis, o espaço é tridimensional. Embora a visualização do espaço não seja possível, podemos ainda rotacionar os eixos buscando maximizar variância da mesma forma que no caso bidimensional.

Após identificarmos a linha na qual a variância é máxima, ainda resta alguma variância em torno da linha, como pode ser observado na Fig. 4.20. Na análise dos componentes principais, após extrairmos o primeiro fator, isto é, após a primeira reta ter sido traçada através dos dados, podemos continuar e definir outra reta que maximiza a variância restante e assim sucessivamente. Uma vez que fatores consecutivos são definidos através da maximização da variabilidade que não é capturada pelo fator precedente, concluímos que os fatores são independentes um do outro. Em outras palavras, fatores consecutivos não são correlacionados, ou seja, são ortogonais entre si.

No exposto acima, utilizamos a técnica de análise dos componentes principais como um método de redução de dados, isto é, um método para reduzir o número de variáveis. Surge portanto a questão de quantos fatores devem

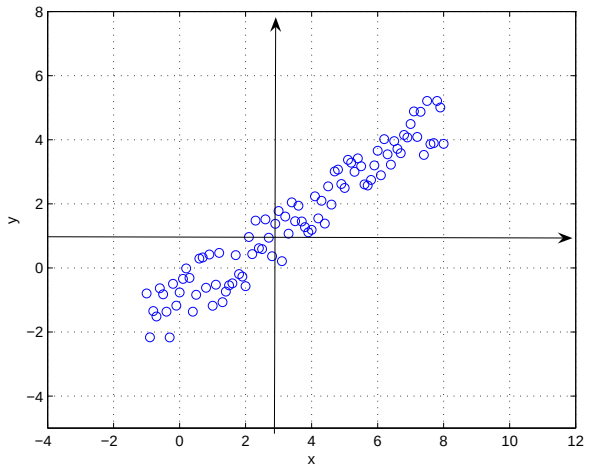


Figura 4.21: Translação do centróide para a origem.

ser extraídos. À medida que extraímos fatores, eles representam cada vez menos a variabilidade observada nos dados. A decisão de quando parar vai depender de quanto a variabilidade residual pode ser atribuída ao fenômeno aleatório, a qual não pode ser capturada por uma relação determinística. A natureza desta decisão é arbitrária e dependente da aplicação. Entretanto, existem linhas gerais e critérios para apoiar nesta tomada de decisão.

Mas o que a decomposição SVD tem a ver com a análise dos componentes principais? A decomposição SVD nos permite computar os componentes principais e também uma medida de quanta variabilidade é capturada por cada um dos componentes. Suponha que nos é dado um conjunto $\{p_1, \ldots, p_n\}$ de vetores correspondendo aos pontos amostrais, onde $p_j \in \mathbb{R}^{m \times 1}$ é um vetor coluna. Montamos a matriz X com os pontos amostrais:

$$X = \begin{bmatrix} | & | & & | \\ p_1 & p_2 & \ldots & p_n \\ | & | & & | \end{bmatrix}$$

Os componentes principais (ou eixos) são os autovetores da matriz XX^T, que podem ser calculados através de uma fatoração por autovalores:

$$XX^T = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} U^T$$

Note que os autovetores de XX^T são precisamente as colunas da matriz U.

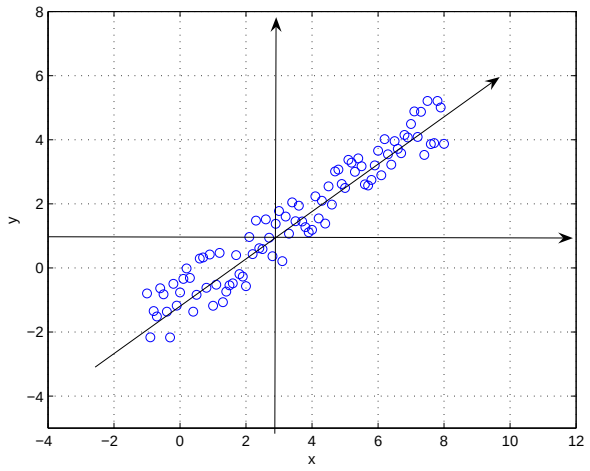


Figura 4.22: Rotação do eixo x em torno da origem.

Podemos computar os componentes principais através da fatoração SVD de X, fazendo:

$$\begin{aligned} X &= U\Sigma V^T \\ XX^T &= U\Sigma V^T (U\Sigma V^T)^T \\ &= U\Sigma V^T V\Sigma^T U^T \\ &= U\Sigma^2 U^T \end{aligned}$$

Logo, os vetores singulares esquerdos (as colunas de U) correspondem aos componentes principais, que são ordenados de acordo com os valores singulares de X (σ_i).

Como exemplo, considere o conjunto de pontos do $\mathbb{R}^3$ dados na Tab. 4.1. Primeiramente, obtemos a matriz:

$$XX^T = \begin{bmatrix} 2025.4 & -21.4 & 1509.4 \\ -21.4 & 101.9 & 491.5 \\ 1509.4 & 491.5 & 3659.3 \end{bmatrix}$$

Realizando a decomposição SVD, produzimos:

$$XX^T = U\Sigma^2U^T$$

$$U = \begin{bmatrix} -0.5036 & 0.8500 & -0.1548 \\ -0.0915 & -0.2307 & -0.9687 \\ -0.8591 & -0.4736 & 0.1940 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4.5965 \times 10^3 & & \\ & 1.1900 \times 10^3 & \\ & & 0 \end{bmatrix}$$

Observe que o primeiro componente captura aproximadamente 79.43% da variabilidade, enquanto o segundo componente captura cerca de 20.57% da variabilidade. Já o terceiro componente não captura nenhuma variabilidade. Isto está de acordo com o experimento realizado, para o qual a dimensão x teve um componente aleatório com fator 5, enquanto a dimensão y teve um componente aleatório com fator 8[1]. Por outro lado, a dimensão z não teve nenhuma aleatoriedade sendo definida pela expressão $z = 0.7x + 5y - 1$. A Fig. 4.23 ilustra os pontos dados na Tab. 4.1. Note que os pontos variam nas dimensões x e y de forma arbitrária, por outro lado, observe que o valor de z é uma função linear de x e y. Em outras palavras, os pontos dados estão contidos em um plano.

4.15 Métodos Iterativos

Os métodos diretos não são os mais indicados para resolver sistemas esparsos e os sistemas de grande porte. O refinamento do algoritmo de Gauss é uma forma de método iterativo. Os algoritmos iterativos partem de uma solução inicial e sistematicamente geram uma sequência de iterandos. Dada uma aproximação inicial x_0 da solução de $Ax = b$, o processo iterativo pode ser descrito como:

$$x_{k+1} = G(x_k),\ k = 1, 2, \ldots,$$

Aqui vamos estudar os métodos de Jacobi e de Gauss-Seidel.

[1]Os pontos (x, y, z) foram gerados através das expressões: $x = 5 * rand() - 12.5)$; $y = 8 * rand() - 4$; $z = 0.7 * x + 5 * y - 1$, onde $rand()$ produz um número aleatório uniformemente distribuído no intervalo $(0, 1)$.

Tabela 4.1: Conjunto de pontos de um experimento

j	x_j	y_j	z_j
1	-7.7494	-2.1509	-17.1790
2	-9.4658	-0.1121	-8.1868
3	-8.0435	2.0968	3.8534
4	-10.2177	-3.8520	-27.4122
5	-8.3930	-0.4424	-9.0869
6	-9.4228	2.3355	4.0815
7	-7.8909	1.9057	3.0046
8	-11.6187	-0.7544	-12.9048
9	-7.8227	3.3352	10.2003
10	-10.4486	3.1492	7.4319
11	-12.2105	-1.1771	-15.4327
12	-8.4342	-3.9211	-26.5095
13	-11.8055	-2.3779	-21.1533
14	-11.5064	0.8303	-4.9028
15	-11.1391	-2.4095	-20.8448
16	-12.4236	1.9743	0.1749
17	-10.2745	3.4545	9.0804
18	-10.1700	-0.6508	-11.3730
19	-8.2689	0.2012	-5.7821
20	-11.4868	1.3771	-2.1552

4.15.1 Método de Jacobi

Seja $Ax = b$ um sistema de equações lineares quadrado, expresso na forma:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{4.39}$$

Assumindo que $a_{jj} \neq 0$ para $j = 1, \dots, n$, podemos expressar (4.39) na forma:

$$\begin{aligned} x_j \;=\; & \frac{1}{a_{jj}} (b_j - a_{j1}x_1 - a_{j2}x_2 - \dots \\ & \dots - a_{j,(j-1)}x_{j-1} - a_{j,(j+1)}x_{j+1} - \dots - a_{j,n}x_n) \end{aligned} \tag{4.40}$$

para $j = 1, \dots, n$, onde são conhecidos os valores de $x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_n$ a serem utilizados no lado direito da expressão acima.

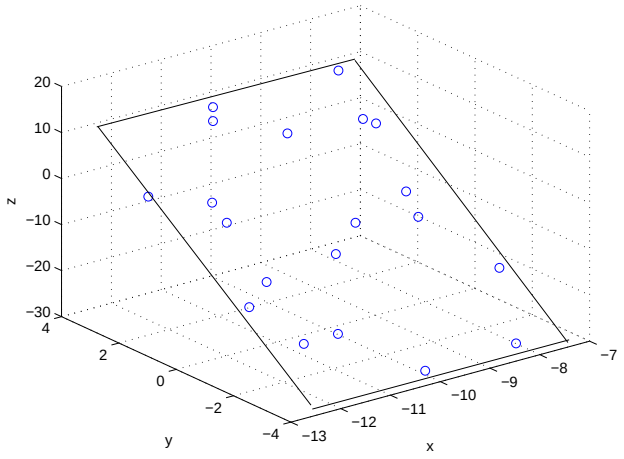


Figura 4.23: Pontos amostrais para análise de componentes principais.

O *Método de Jacobi* consiste no processo iterativo dado por

$$x_j^{k+1} = \frac{1}{a_{jj}}\left(b_j - a_{j1}x_1^k - a_{j2}x_2^k - \ldots - a_{j,(j-1)}x_{j-1}^k - a_{j,(j+1)}x_{j+1}^k - \ldots - a_{j,n}x_n^k\right), j = 1, \ldots, n \quad (4.41)$$

Se o sistema $Ax = y$ é não singular, sempre será possível obter $a_{jj} \neq 0$ se trocarmos as equações adequadamente. Em sistemas esparsos, o método de Jacobi evita multiplicações desnecessárias, simplificando a expressão acima.

Jacobi$(x_1, \ldots, x_n, a_{ij}, b_j, lim)$

$k \leftarrow 1$
while $k < lim$ **do**
  for $j = 1, \ldots, n$ **do**
    $z_j = \frac{1}{a_{jj}}(b_j - a_{j1}x_1 - a_{j2}x_2 - \ldots - a_{j,(j-1)}x_{j-1} - a_{j,(j+1)}x_{j+1} - \ldots - a_{j,n}x_n)$
  end for
  if Teste de convergência é satisfeito **then**
    Saída $(z_j : j = 1, \ldots, n)$
  else
    $x_j \leftarrow z_j$, $j = 1, \ldots, n$
  end if
  $k \leftarrow k + 1$
end while
Saída "Algoritmo não convergente"

Exemplo do Método de Jacobi

Vamos tomar como exemplo o sistema de equações lineares dado por:

$$\begin{cases} 5x_1 + & & & & - 3x_4 - x_5 = 2 \\ -x_1 + 4x_2 & & & & - x_5 = 3 \\ & & + 2x_3 - x_4 & & = -1 \\ -x_1 & & & & + 4x_4 - 2x_5 = 0 \\ & & & & - x_4 + 2x_5 = -1 \end{cases} \tag{4.42}$$

Primeiramente, obtemos o operador iterativo que corresponde às equações:

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(2 + 3x_4^k + x_5^k) \\ x_2^{k+1} = \frac{1}{4}(3 + x_1^k + x_5^k) \\ x_3^{k+1} = \frac{1}{2}(-1 + x_4^k) \\ x_4^{k+1} = \frac{1}{4}(x_1^k + 2x_5^k) \\ x_5^{k+1} = \frac{1}{2}(-1 + x_4^k) \end{cases} \tag{4.43}$$

O processo iterativo acima ao ser aplicado na resolução de (4.42) produz a seqüência de iterandos dada na Tabela 4.2.

Tabela 4.2: iterandos obtidos pelo processo iterativo de Jacobi

Iteração	1	2	...	35	36
x_1	1	1.20	...	0.0869574059	0.086957108
x_2	1	1.25	...	0.6086962107	0.608696020
x_3	1	0.00	...	-0.6521793252	-0.652173522
x_4	1	0.75	...	-0.3043470441	-0.304347311
x_5	1	0.00	...	-0.6521733252	-0.652173522

4.15.2 Método de Gauss-Seidel

No método anterior observamos que, quando x_2^{k+1} é calculado, na realidade já possuímos o valor x_1^{k+1} que é mais próximo da solução do que x_1^k. Isto sugere o método Gauss-Seidel, que corresponde ao processo iterativo modificado:

$$\begin{aligned} x_j^{k+1} = \frac{1}{a_{ii}}(b_j - a_{j1}x_1^{k+1} - a_{j2}x_2^{k+1} - \ldots - a_{j,(j-1)}x_{j-1}^{k+1} \\ -a_{j,(j+1)}x_{j+1}^k - a_{j,n}x_n^k), \quad j = 1, \ldots, n \end{aligned} \tag{4.44}$$

Gauss_Seidel$(x_1, \ldots, x_n, a_{ij}, b_j, lim)$

1: $k \leftarrow 1$
2: **while** $k < lim$ **do**
3: **for** $j = 1, \ldots, n$ **do**
4: $z_j = \frac{1}{a_{jj}}(b_j - a_{j1}z_1 - a_{j2}z_2 - \ldots - a_{j,(j-1)}z_{j-1} - a_{j,(j+1)}x_{j+1} - \ldots - a_{j,n}x_n)$
5: **end for**
6: **if** Teste de convergência é satisfeito **then**
7: Saída $(z_j : j = 1, \ldots, n)$
8: **else**
9: $x_j \leftarrow z_j, j = 1, \ldots, n$
10: **end if**
11: $k \leftarrow k + 1$
12: **end while**
13: Saída "Algoritmo não convergente"

Observação: se o teste de convergência não usa os valores x_{j-1} então o processo iterativo pode ser substituído por:

$$\begin{aligned} x_j &= \frac{1}{a_{jj}}(b_j - a_{j1}x_1 - a_{j2}x_2 - \ldots \\ &\qquad -a_{j,(j-1)}x_{j-1} - a_{j,(j+1)}x_{j+1} - \ldots - a_{j,n}x_n), \ j = 1, \ldots, n \end{aligned}$$

Resultados do algoritmo ao ser aplicado ao exemplo anterior, resolução do sistema de equações lineares (4.42), é ilustrado na Tabela 4.3.

Tabela 4.3: Iterandos obtidos pelo processo iterativo de Gauss-Seidel

Iteração	1	2	3	...	18
x_1	1	1.200	0.8600	...	0.0869585
x_2	1	1.300	0.9400	...	0.6086963
x_3	1	0.000	-0.1000	...	-0.6521724
x_4	1	0.800	0.1650	...	-0.3043465
x_5	1	-0.100	-0.4175	...	-0.6521732

4.15.3 Visão Geométrica do Método Iterativo

Para fins de ilustração, vamos considerar um sistema linear genérico em duas variáveis em duas equações:

$$\begin{cases} ax + by = m \\ cx + dy = n \end{cases}$$

portanto,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \text{ e } b = \begin{bmatrix} m \\ n \end{bmatrix}$$

Processo iterativo do método de Jacobi é dado por:

$$\begin{cases} x_{k+1} &= \frac{1}{a}(m - by_k) \\ y_{k+1} &= \frac{1}{d}(n - cx_k) \end{cases}$$

enquanto que o processo iterativo do método de Gauss-Seidel toma a forma:

$$\begin{cases} x_{k+1} = \frac{1}{a}(m - by_k) \\ y_{k+1} = \frac{1}{d}(n - cx_{k+1}) \end{cases}$$

As Figuras 4.24 e 4.25 ilustram graficamente o comportamento do processo iterativo Gauss-Seidel para os casos de convergência e divergência, respectivamente.

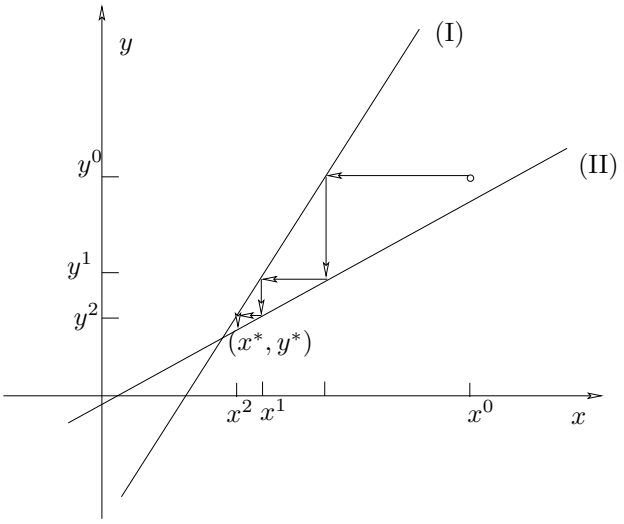


Figura 4.24: Visão geométrica do processo iterativo de Gauss-Seidel, caso convergente.

Algoritmos Preparadores para o Gauss-Seidel/Jacobi:

Objetivo de um algoritmo preparador é transformar o sistema $Ax = b$ em uma forma equivalente $x = Bx + d$, de modo a se obter um esquema iterativo convergente expresso na forma $x^{k+1} = Bx^k + d$. Dois preparadores são o *preparador clássico* e o *preparador unitário*.

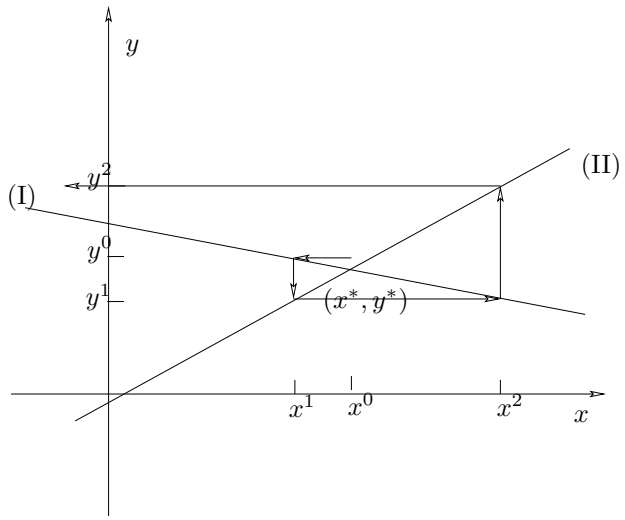


Figura 4.25: Visão geométrica do processo iterativo de Gauss-Seidel, caso divergente.

1) *Preparador Clássico:*

$$\begin{cases} x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n) \\ x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3 - \ldots - a_{2n}x_n) \\ \vdots &= \qquad\qquad \vdots \\ x_n &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{n,(n-1)}x_{n-1}) \end{cases}$$

o qual pode ser escrito de uma forma mais compacta

$$x_j = \frac{1}{a_{jj}}\left(b_j - \sum_{k=1,k\neq j}^{n} a_{jk}x_k\right),\ j = 1, \ldots, n$$

2) *Preparador Unitário:*

$$\begin{cases} x_1 &= b_1 + (1 - a_{11})x_1 - a_{12}x_2 - \ldots - a_{1n}x_n \\ x_2 &= b_2 - a_{21}x_1 + (1 - a_{22})x_2 - \ldots - a_{2n}x_n \\ \vdots &= \qquad\qquad \vdots \\ x_n &= b_n - a_{n1}x_1 - a_{n2}x_2 - \ldots + (1 - a_{nn})x_n \end{cases}$$

que também pode ser colocado de forma mais compacta:

$$x_j = b_j - \sum_{k=1}^{j-1} a_{jk}x_k + (1 - a_{jj})x_j - \sum_{k=j+1}^{n} a_{jk}x_k,\ j = 1, \ldots, n$$

4.15.4 Condições de Convergência dos Métodos Iterativos

As condições de convergência estão relacionadas com os algoritmos preparadores. Vamos inicialmente estudar a condição de convergência do método de Jacobi para o preparador clássico.

Teorema 4.10 *Seja o sistema de equações lineares modificado $x = Bx + d$. Se $\|B\| < 1$, então o algoritmo de Jacobi converge.*

Prova: Seja x' é uma solução de $Ax = b \Leftrightarrow x = Bx + d$. Então,

$$\begin{aligned} \|x^{k+1} - x'\| &= \|Bx^k + d - x'\| \\ &= \|Bx^k + d - (Bx' + d)\| \\ &= \|B(x^k - x')\| \\ &\leqslant \|B\| \cdot \|x^k - x'\| \\ &< \|x^k - x'\| \end{aligned}$$

Portanto $\lim_{k\to\infty} \|x^k - x'\| = 0$. ■

Corolário 4.1 *Seja o sistema $Ax = y$. Então, se a matriz A é diagonalmente dominante, $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$ para $i = 1, \ldots, n$, então o método de Jacobi é convergente.*

Como exemplo, considere o operador iterativo de Jacobi dado por (4.43). Este operador pode ser escrito na forma $x^{(k+1)} = Bx^{(k)} + d$, onde:

$$B = \begin{bmatrix} 0 & 0 & 0 & 3/5 & 1/5 \\ 1/4 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 & 0 \\ 1/4 & 0 & 0 & 0 & 2/4 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \qquad b = \begin{bmatrix} 2/5 \\ 3/4 \\ -1/2 \\ 0 \\ -1/2 \end{bmatrix}$$

Note que $\|B\|_\infty = 0.8$, o que implica a convergência do operador iterativo.

4.16 Método do Gradiente Conjugado

Aqui, consideramos o problema de resolver o sistema de equações lineares:

$$Ax = b \tag{4.45}$$

onde $A = A^T \in \mathbb{R}^{n\times n}$ é uma matriz positiva definida.

Existe uma relação entre a solução do sistema $Ax = b$ e o problema de encontrar o mínimo de uma função quadrática:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^T Ax - b^T x \tag{4.46}$$

Uma vez que f é convexa, o ponto mínimo $x^\star$ de (4.46) induz gradiente de f nulo:

$$\nabla f(x^\star) = Ax^\star - b = 0$$

Isto sugere que o sistema linear (4.45) pode ser resolvido através de métodos iterativos de otimização baseados em gradiente. A partir de um chute inicial $x^{(0)}$, tais métodos geram uma sequência $\{x^{(k)}\}$ que converge para a solução $x^\star$. Dado o iterando $x^{(k)}$, o método de descenso produz $x^{(k+1)}$ fazendo:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k \nabla f(x^{(k)}) \\ &= x^{(k)} + \alpha_k (b - Ax^{(k)}) \\ &= x^{(k)} + \alpha_k p_k \end{aligned}$$

onde p_k é a direção de descenso que corresponde ao negativo do gradiente de f no ponto $x^{(k)}$. O vetor de resíduos $r_k = b - Ax^{(k)}$ define a direção de descenso. A escolha do passo α_k é fundamental para convergência global do método de descenso. O método do gradiente conjugado consiste de um método de descenso onde as direções p_k são mutuamente conjugadas. Ou seja, qualquer par de direções p_i e p_j da sequência $\{p_k\}$ satisfaz a relação:

$$p_i^T A p_j = 0$$

Duas direções p_i e p_j são ditas conjugadas se são ortogonais com respeito ao produto interno $\langle A^T p_i, p_j\rangle = \langle p_i, Ap_j\rangle = p_i^T Ap_j$.

Note que um conjunto $\{p_k\}$ com n direções mutuamente conjugadas forma uma base do $\mathbb{R}^n$. Logo, a solução $x^\star$ para $Ax = b$ pode ser escrita da seguinte forma:

$$x^\star = \alpha_1 p_1 + \alpha_2 p_2 + \cdots + \alpha_n p_n \tag{4.47}$$

Os coeficientes α_k são dados por:

$$Ax^\star = \alpha_1 Ap_1 + \alpha_2 Ap_2 + \cdots + \alpha_n Ap_n \tag{4.48}$$

$$\begin{aligned} p_k^T Ax^\star &= p_k^T \alpha_1 Ap_1 + p_k^T \alpha_2 Ap_2 + \cdots + p_k^T \alpha_n Ap_n \\ &= \alpha_k p_k^T Ap_k \end{aligned} \tag{4.49}$$

$$\alpha_k = \frac{p_k^T Ax^\star}{p_k^T Ap_k} = \frac{p_k^T b}{p_k^T Ap_k} \tag{4.50}$$

Portanto, de posse de um conjunto $\{p_k\}$ de n direções mutuamente conjugadas, a solução $x^\star$ pode ser facilmente obtida através dos coeficientes α_k calculados com a expressão (4.50) que são substituídos em (4.47).

O método do gradiente conjugado produz o conjunto $\{p_k\}$ através dos gradientes obtidos por meio dos resíduos r_k em um processo iterativo. Seja $x^{(0)}$ o iterando inicial e defina $r^{(0)} = b - Ax^{(0)}$ e $p_0 = -\nabla f(x^{(0)}) = r_0$. O conjunto $\{p_k\}_{k=0}^{n-1}$ é obtido como segue:

$$p_{k+1} = r_k - \frac{p_k^T A r_k}{p_k^T A p_k} p_k, \quad k = 0, \ldots, n-2 \tag{4.51}$$

Portanto, chegamos ao algoritmo do gradiente conjugado (GC) conforme pseudo-código abaixo:

Gradiente-Conjugado$(A, b, x^{(0)}, \epsilon, n)$

$r_0 \leftarrow b - Ax^{(0)}$
$p_0 = r_0$
$k \leftarrow 0$
while $\|r_k\| > \epsilon$ **do**
  $\alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k}$
  $x^{(k+1)} \leftarrow x^{(k)} + \alpha_k p^{(k)}$
  $r_{k+1} = r_k - \alpha_k A p_k$
  $\beta_k \leftarrow -\frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  $p_{k+1} \leftarrow r_{k+1} + \beta_k p_k$
  $k \leftarrow k + 1$
end while

O método do gradiente conjugado termina em no máximo n passsos se erros de arredondamento não são encontrados. Tipicamente, a solução $x^{(n)}$ é uma excelente aproximação. O número de passos não deve exceder substancialmente em n passos. É preferível interromper o processo iterativo e reiniciá-lo a partir da solução aproximada encontrada na primeira etapa.

O método do gradiente conjugado pode ser aplicado quando A é uma matriz arbitrária resolvendo o sistema normal:

$$A^T A x = A^T b \tag{4.52}$$

uma vez que $A^T A$ é simétrica e positiva semi-definida qualquer que seja a matriz A. Sendo um sistema iterativo, não é necessário montar $A^T A$ de forma explícita em memória, mas simplesmente executar operações matriciais e vetoriais embutidas no operador iterativo. Portanto, o método GC é

particularmente eficiente quando A é uma matriz esparsa. As seguintes modificações são suficientes para que o algoritmo Gradiente-Conjugado aceite uma matriz A qualquer:

$$\begin{aligned} p_0 &= A^T r_0 \\ \alpha_k &= \frac{r_k^T A A^T r_k}{p_k^T A^T A p_k} \\ \beta_k &= \frac{r_{k+1}^T A A^T r_{k+1}}{r_k^T A A^T r_k} \\ p_{k+1} &= A^T r_{k+1} + \beta_k p_k \end{aligned}$$

Porém, a desvantagem em montar as equações normais (4.52) é que o número de condicionamento $\kappa(A^T A)$ é igual a $\kappa(A)^2$, assim reduzindo a taxa de convergência do método GC.

Aplicando-se o método do gradiente conjugado ao sistema linear (4.42), obtém-se os iterandos dados na Tabela 4.4. As direções conjugadas com respeito à matriz $A^T A$ são dadas na Tabela 4.5. Note $\langle p^{(i)}, A^T A p^{(j)} \rangle = 0$ para $i \neq j$.

Tabela 4.4: iterandos produzidos pelo método do gradiente conjugado

$x^{(k)} \backslash k$	0	1	2	3	4	5
$x_1^{(k)}$	62.9447	71.8333	42.8618	28.8714	30.1741	0.0870
$x_2^{(k)}$	81.1584	48.8943	24.0198	25.0221	17.7009	0.6087
$x_3^{(k)}$	-74.6026	-58.5635	-37.5481	19.1819	21.6932	-0.6522
$x_4^{(k)}$	82.6752	47.7951	48.5802	32.1188	33.4815	-0.3043
$x_5^{(k)}$	26.4718	52.7845	69.3574	44.9740	37.5150	-0.6522

Tabela 4.5: direções conjugadas

$p^{(k)} \backslash k$	0	1	2	3	4
$p_1^{(k)}$	0.2559e+3	-645.4612	-75.3341	16.3974	-23.4300
$p_2^{(k)}$	-0.9289e+3	-554.1837	5.3973	-92.1580	-13.3103
$p_3^{(k)}$	0.4618e+3	468.2044	305.4747	31.6107	-17.4012
$p_4^{(k)}$	-1.0042e+3	17.4912	-88.6395	17.1533	-26.3103
$p_5^{(k)}$	0.7575e+3	369.2289	-131.2972	-93.8918	-29.7223

4.17 Aplicações de Sistemas de Equações Lineares

Nesta seção ilustramos aplicações de sistemas de equações lineares a problemas de interesse.

4.17.1 Interpolação com Polinômios

Problema: dados os pontos $D = \{(t_1, y_1), \ldots, (t_n, y_n)\}$ construa um polinômio $p(t) = x_1 + x_2 t + x_3 t^2 + \ldots + x_n t^{n-1}$ que passe pelos pontos dados. Portanto, os dados são os pontos $\{(t_1, y_1), \ldots, (t_n, y_n)\}$ e as variáveis são $\{x_1, x_2, \ldots, x_n\}$, que correspondem aos coeficientes do polinômio. A Figura 4.26 ilustra uma função desconhecida $f(x)$, para a qual conhecemos o seu valor em alguns pontos e desejamos encontrar uma interpolação.

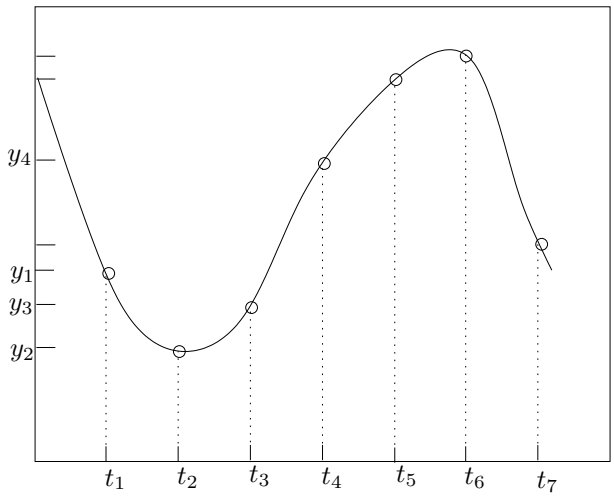


Figura 4.26: Exemplo de dados para interpolação com polinômios.

Podemos transformar o problema na solução de um sistema de equações lineares:

$$\left\{\begin{array}{lcccccccccc} p(t_1) & = & x_1 & + & x_2 t_1 & + & x_3 t_1^2 & + & \ldots & + & x_n t_1^{n-1} & = & y_1 \\ p(t_2) & = & x_2 & + & x_2 t_2 & + & x_3 t_2^2 & + & \ldots & + & x_n t_2^{n-1} & = & y_2 \\ \vdots & = & \vdots & & & & & & & & & & \\ p(t_n) & = & x_n & + & x_2 t_n & + & x_3 t_n^2 & + & \ldots & + & x_n t_n^{n-1} & = & y_n \end{array}\right. \tag{4.53}$$

O sistema (4.53) pode ser colocado na forma matricial:

$$\begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \dots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & \dots & t_n^{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Portanto, a busca do polinômio interpolador $p(t)$ para o conjunto de pontos D, i.e. um polinômio que passe pelos elementos de D, foi reduzida à solução de um sistema de equações lineares.

4.17.2 Estruturas Elásticas Lineares

Considere uma estrutura conforme Figura 4.27, sendo esta formada em geral por n nós (pontos de junção conectados por barras) e um conjunto de forças externas aplicadas aos pontos dado por $f \in \mathbb{R}^{2n}$ (vetor de forças externas, vertical e horizontal, aplicadas aos pontos). Nosso problema é determinar os alongamentos/compressões provocados nas barras pelas cargas, ou seja, queremos determinar o vetor de alongamentos verticais e horizontais $\Delta s \in \mathbb{R}^{2n}$.

Para $\|\Delta s\|$ pequeno, o deslocamento pode ser aproximado por uma função linear $f = A\Delta s$ onde A é dita matriz de resistência a qual depende do material e da geometria. Portanto, o problema de encontrar os alongamentos se reduz a obter a matriz A e resolver o sistema de equações $f = A\Delta s$, sendo o vetor de forças externas f (cargas) conhecido.

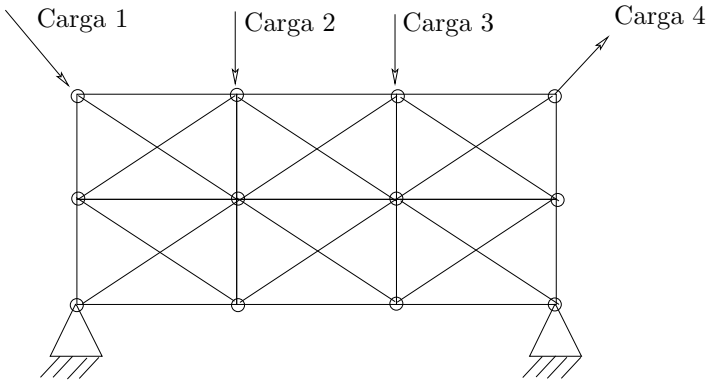


Figura 4.27: Exemplo de estrutura.

4.17.3 Exemplo: n=1

Aqui ilustramos o processo para o caso mais simples, onde o número de vértices é $n = 1$, conforme a ilustração dada na Figura 4.28.

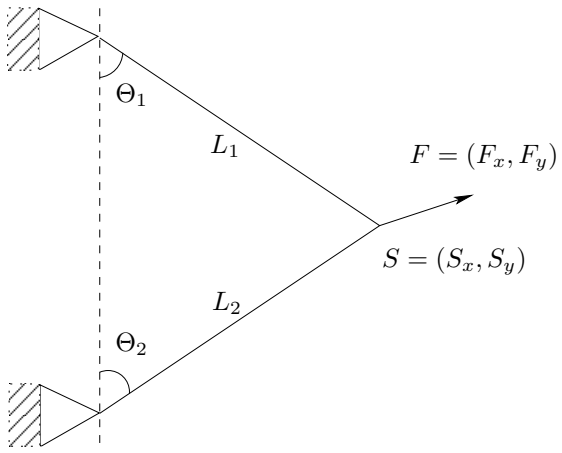


Figura 4.28: Exemplo de estrutura com apenas um vértice

Quando as cargas aplicadas são nulas, $f = 0$, a posição do nó é $s = (s_x, s_y)$ como indicado na figura. Quando uma carga é aplicada, $f \neq 0$, a posição do nó é dada por $s + \Delta s = (s_x + \Delta s_x, s_y + \Delta s_y)$. No que segue vamos mostrar que para $\|\Delta s\|$ pequeno, $\|\Delta s\|$ é dado por

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \Delta s_x \\ \Delta s_y \end{bmatrix} \Rightarrow f = A\Delta s$$

onde os coeficientes a_{ij} dependem apenas do material e da geometria da estrutura. Considere agora a Figura 4.29. Assumindo que o nó superior está localizado na origem, deduzimos que

$$\begin{aligned} s_x &= l_1 \sin\theta_1 \\ s_y &= -l_1 \cos\theta_1 \\ l_1 + \Delta l_1 &= \sqrt{(s_x + \Delta s_x)^2 + (s_y + \Delta s_y)^2} \\ &\approx \sqrt{s_x^2 + s_y^2} + \frac{s_x \Delta s_x + s_y \Delta s_y}{\sqrt{s_x^2 + s_y^2}} \end{aligned}$$

A aproximação pode ser obtida através da expansão de Taylor de primeira ordem. Por exemplo, fazendo $g(s_x, s_y) = \sqrt{s_x^2 + s_y^2}$, $s'_x = s_x + \Delta s_x$, e $s'_y =$

$s_y + \Delta s_y$, podemos calcular

$$\begin{aligned} g(s'_x, s'_y) &\approx g(s_x, s_y) + \frac{\partial g(s_x, s_y)}{\partial s_x}(s'_x - s_x) + \frac{\partial g(s_x, s_y)}{\partial s_y}(s'_y - s_y) \\ &= g(s_x, s_y) + \frac{\partial g(s_x, s_y)}{\partial s_x}\Delta s_x + \frac{\partial g(s_x, s_y)}{\partial s_y}\Delta s_y \\ &= \sqrt{s_x^2 + s_y^2} + \frac{1}{2}\frac{2s_x}{\sqrt{s_x^2 + s_y^2}}\Delta s_x + \frac{1}{2}\frac{2s_y}{\sqrt{s_x^2 + s_y^2}}\Delta s_y \\ &= \sqrt{s_x^2 + s_y^2} + \frac{s_x\Delta s_x + s_y\Delta s_y}{\sqrt{s_x^2 + s_y^2}} \end{aligned}$$

Portanto,

$$l_1 + \Delta l_1 = l_1 + \sin\theta_1\Delta s_x - \cos\theta_1\Delta s_y$$

e para, Δs pequeno, temos que $\Delta l_1 = \sin\theta_1\Delta s_x - \cos\theta_1\Delta s_y$.

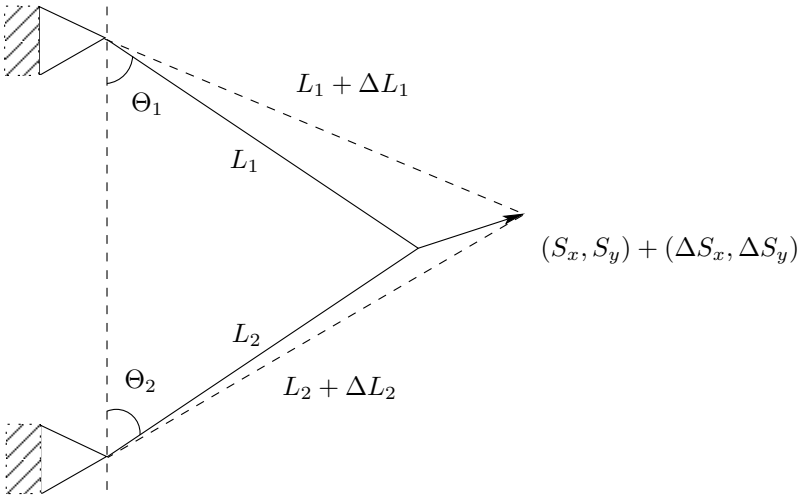


Figura 4.29: Exemplo de estrutura com apenas um vértice e carga não nula

De maneira similar ao desenvolvimento acima, podemos concluir que

$$\Delta l_2 = \sin\theta_2\Delta s_x + \cos\theta_2\Delta s_y$$

Em notação matricial, o problema fica:

$$\begin{bmatrix} \Delta l_1 \\ \Delta l_2 \end{bmatrix} = \begin{bmatrix} \sin\theta_1 & -\cos\theta_1 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} \Delta s_x \\ \Delta s_y \end{bmatrix}$$

Vamos agora considerar as forças necessárias para deformar barras (material elástico), conforme ilustração nas Figuras 4.30 e 4.31. Mais precisamente,

- F_1 e F_2 são forças internas às barras;
- k_1 e k_2 são constantes que dependem da geometria e do material.

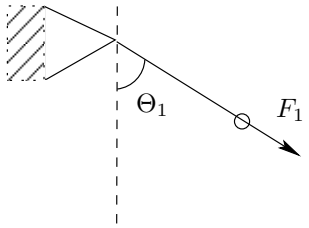


Figura 4.30: Forças internas às barras

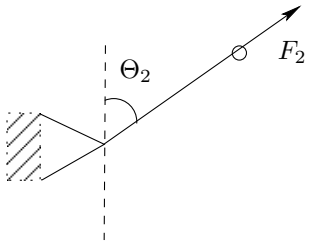


Figura 4.31: Forças internas às barras

O próximo passo consiste do cálculo das forças de equilíbrio nos nós. As forças que atuam no nó são ilustradas na Figura 4.32. Decompondo as forças internas as barras, as quais devem equilibrar as forças externas, obtemos as equações:

$$\begin{aligned} F_1 \sin\theta_1 + F_2 \sin\theta_2 &= f_x \\ -F_1 \cos\theta_1 + F_2 \cos\theta_2 &= f_y \end{aligned}$$

Sabemos que

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} \Delta l_1 \\ \Delta l_2 \end{bmatrix}$$

e

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \sin\theta_1 & \sin\theta_2 \\ -\cos\theta_1 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}$$

Portanto

$$\begin{aligned} \begin{bmatrix} f_x \\ f_y \end{bmatrix} &= \begin{bmatrix} \sin\theta_1 & \sin\theta_2 \\ -\cos\theta_1 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} \sin\theta_1 & -\cos\theta_1 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} \Delta s_x \\ \Delta s_y \end{bmatrix} \\ &= A \begin{bmatrix} \Delta s_x \\ \Delta s_y \end{bmatrix} \end{aligned}$$

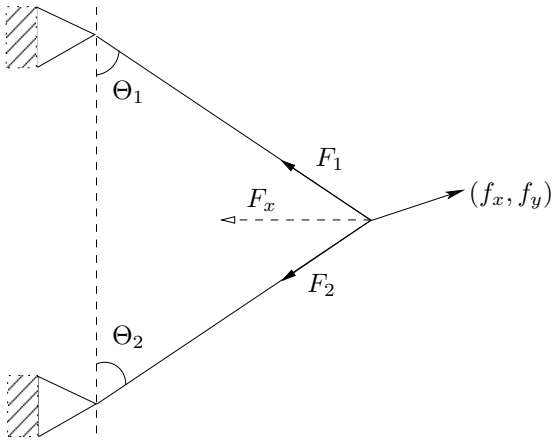


Figura 4.32: Forças que atuam no nó

4.18 Referências

O texto deste capítulo é uma síntese do texto de Cláudio and Marins [1]. Em particular, as seguintes seções são resumos de [1] com modificações e introdução de exemplos:

- seção 4.2 (erros computacionais);
- seção 4.3 (etapas de solução de sistemas lineares);
- seção 4.4 (método de eliminação de Gauss);
- seção 4.5 (instabilidade numérica);
- seção 4.6 (método de Gauss com pivoteamento);
- seção 4.7 (condicionamento de matrizes);
- seção 4.8 (refinamento de solução por meio do método de Gauss);
- seção 4.9 (equacionamento matricial);
- seção 4.15 (métodos iterativos).

O exemplo de aplicação em estruturas elásticas lineares foi retirado das notas de aula da disciplina *EE103: Applied Numerical Computing* [7]. As seções sobre decomposição LU e SVD foram inspiradas nos textos de Press et al. [5] e Cormen et al. [2].

4.19 Excercícios

Exercício 4.1 Seja a equação polinomial abaixo, chamada *equação Diofantina*, utilizada na teoria de controle:

$$A(s)D(s) + B(s)N(s) = \Delta(s) \tag{4.54}$$

Nesta equação são conhecidos os polinômios $N(s)$ e $D(s)$ e deseja-se determinar $A(s)$ e $B(s)$. Para tanto, considera-se também como conhecido o polinômio característico desejado, $\Delta(s)$, o qual deve ter todas as suas raízes com parte real negativa: $Re(s_i) < 0$. Nos itens seguintes, utilizar-se-á:

$$\begin{aligned} N(s) &= n_0 + n_1 s + n_2 s^2 \\ D(s) &= d_0 + d_1 s + d_2 s^2 \\ \Delta(s) &= \delta_0 + \delta_1 s + \delta_2 s^2 + \delta_3 s^3 \\ A(s) &= a_0 + a_1 s \\ B(s) &= b_0 + b_1 s \end{aligned}$$

sendo $d_2 \neq 0$. Tarefas:

a) Mostre que a solução da *equação Diofantina* (4.54) pode ser obtida através da solução do sistema de equações lineares (4.55), cujas incógnitas são os coeficientes dos polinômios $A(s)$ e $B(s)$:

$$\begin{bmatrix} d_0 & 0 & n_0 & 0 \\ d_1 & d_0 & n_1 & n_0 \\ d_2 & d_1 & n_2 & n_1 \\ 0 & d_2 & 0 & n_2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{4.55}$$

b) Considere $N(s) = 1 + s$ e $D(s) = 1 + 2.5s + s^2$ e $\Delta(s) = (4 + s)(10 + 6s + s^2) = 40 + 34s + 10s^2 + s^3$. Monte o sistema (4.55) correspondente e utilize o método de Gauss com pivoteamento para determinar o sistema triangular que permite determinar os coeficientes de $A(s)$ e $B(s)$. (Mostrar a cada passo as trocas de linhas realizadas, o pivô e os os fatores multiplicativos).

c) Sejam $N(s)$, $D(s)$ e $\Delta(s)$ utilizados no item anterior. Considere $A(s) = s$ e reformule o sistema (4.55) para encontrar os coeficientes de $B(s)$ utilizando o método dos mínimos quadrados. Dê o polinômio característico que é efetivamente obtido ao se utilizar $A(s)$ dado e $B(s)$ calculado.

d) Considere a matriz definida positiva $P = S^T S$, obtida a partir de:

$$S^T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Determine a decomposição de Choleski de P.

Exercício 4.2 O sistema de equações lineares a coeficientes reais a seguir foi montado para encontrar a solução $x \in \mathcal{C}^2$ de um sistema de equações a coeficientes complexos $Ax = b$, com $A \in \mathcal{C}^{2\times 2}$ e $b \in \mathcal{C}^2$:

$$\begin{bmatrix} 3 & 0.7 & -1 & 1 \\ 1 & 2 & -1 & 0 \\ 1 & -1 & 3 & 0.7 \\ 1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Para o sistema a coeficientes reais acima:

a) Aplique o critério das linhas e conclua sobre a convergência dos métodos de Jacobi e de Gauss-Siedel.

b) Aplique o critério de Sassenfeld e conclua sobre a convergência do método de Gauss-Siedel.

c) Faça uma iteração do método de Gauss-Siedel; considere a aproximação inicial:

$$\begin{bmatrix} t_{10} & t_{20} & v_{10} & v_{20} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$$

d) Determine o sistema de equações lineares a coeficientes complexos correspondente.

Exercício 4.3 Considere o sistema de equações lineares $Ax = b$ e o processo iterativo $x = Bx + d$ correspondendo ao método de Jacobi. Responda as questões abaixo.

a) Prof. Clóvis afirma que se $\|B\| < 1$ então o método de Jacobi sempre converge, qualquer que seja o ponto inicial $x(0)$. Isto é diferente do método de Newton, cuja convergência depende do ponto inicial. Você concorda com a afirmação do Prof. Clóvis? Justifique a sua resposta.

b) Prof. Clóvis também afirma que se o processo iterativo $x(k+1) = Bx(k) + d$ converge para um ponto fixo x^* a partir de um ponto inicial $x(0) \neq x^*$, então o processo iterativo sempre converge para um ponto fixo qualquer ponto inicial x'. Você concorda ou discorda da afirmação? Justique a sua resposta.

Exercício 4.4 Seja x a solução do sistema linear $Ax = b$, onde $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$. Considere o sistema $Ay = b+\Delta b$ obtido ao perturbarmos o lado direito do sistema original. Prof. Wilson afirma que se conhecermos $\kappa(A)$, o número de condicionamento da matriz A, então pode-se estabelecer uma região $S \subseteq \mathbb{R}^n$ que depende de $\kappa(A)$ tal que $y \in S$. Se você discorda do Prof. Wilson, justifique a resposta. Se você concorda com o Prof. Wilson, mostre como que se obtém S?

Exercício 4.5 Considere o sistema de equações lineares $Ax = b$, onde $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$ com $n = 2$. Ao aplicarmos o preparador clássico do método de Jacobi, obtemos $x = Bx + d$. É possível que o método de Jacobi venha a se comportar como o método de Gauss-Seidel? Para uma resposta negativa, justifique a resposta. Para uma resposta afirmativa, indique que condição leva ao mesmo comportamento dos métodos.

Exercício 4.6 Considere os sistemas abaixo indicados:

$$Ax_1 = b_1,\ Ax_2 = b_2,\ \ldots,\ Ax_k = b_k,$$

onde $A \in \mathbb{R}^{n\times n}$ e $x_j, b_j \in \mathbb{R}^n$ para $j = 1, \ldots, k$. Como que você encontraria as soluções $x_1, \ldots x_k$? O seu método deve ser o mais eficiente possível.

Exercício 4.7 Seja $Ax = b$ um sistema de equações lineares, onde $A \in \mathbb{R}^{n\times n}$ e $x, b \in \mathbb{R}^n$ com $n = 2$. A Profa. Mirian afirma que se $det(A) = 0$, então qualquer que seja o preparador do método de Gauss-Seidel, o método iterativo não converge. Você concorda ou discorda da afirmação? Justifique sua resposta.

Exercício 4.8 Aplique os métodos de Gauss com pivoteamento, método iterativo de Jacobi, método iterativo de Gauss-Seidel, e decomposição LU para resolver o sistema de equações abaixo: (descreva todos os passos executados por cada método)

$$\begin{cases} 2x_1 - 3x_2 + 12x_3 - 5x_4 = 6 \\ 5x_1 + 10x_2 - 3x_3 + 2x_4 = 14 \\ x_1 + 2x_2 - 5x_3 + 12x_4 = 10 \\ 10x_1 + 5x_2 + 2x_3 + x_4 = 18 \end{cases}$$

Exercício 4.9 Seja $\|M\| = \sqrt{\lambda_{max}(M^T M)}$ uma norma-matricial para $M \in \mathbb{R}^{n\times n}$, onde $\lambda_{max}(M)$ é definido como o maior autovalor da matriz M. Considere as matrizes quadradas abaixo:

$$A = \begin{bmatrix} -1/3 & 0 & 0 \\ 23 & 1 & 0 \\ -3/7 & 7/8 & -9 \end{bmatrix} \quad B = \begin{bmatrix} -10 & 1 & 5 \\ 5 & -6 & 7 \\ 8 & 33 & 4 \end{bmatrix} \quad C = \begin{bmatrix} -1/3 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -4 \end{bmatrix}$$

Calcule $\|A\|_\infty$, $\|B\|_1$ e $\|C\|$.

Exercício 4.10 Seja o sistema $Ax = b$ dado pelas equações abaixo:

$$\begin{cases} 3x_1 - \frac{1}{3}x_2 + \frac{1}{4}x_3 = -10.5 \\ x_1 + 2x_2 + x_4 = 6 \\ 2.1x_1 + 2x_2 - 3.5x_3 = 3.7 \end{cases}$$

Prof. Kuhn afirma que se o método de Jacobi for obtido a partir do preparador clássico, então para qualquer ponto inicial $x^0 \in \mathbb{R}^3$ o processo iterativo resultante converge para uma solução. Você concorda ou discorda da afirmação do Prof. Kuhn? Justifique a sua resposta.

Exercício 4.11 As questões abaixo se referem a um sistema $Ax = b$ onde $A \in \mathbb{R}^{m\times n}$. (**Justifique as respostas.**)

i. Se $m = n$, como você faria para calcular o condicionamento de A?

ii. Prof. Kuhn afirma que se $n > m$, ou seja, se há mais variáveis do que equações, então o sistema tem uma ou mais soluções. Você concorda ou discorda da afirmação?

iii. Prof. Kuhn afirma que se o sistema linear tem mais do que uma solução, então devem existir infinitas soluções. Você concorda ou discorda?

iv. Prof. Kuhn também afirma que se $rank(A^T) = n$ e $b \in range(A)$ então o sistema tem solução única. Você concorda ou discorda?

Exercício 4.12 Estamos interessados em encontrar uma solução para o sistema $Ax = b$, onde $A \in \mathbb{R}^{n\times n}$.

i. Sob quais condições o sistema tem um número infinito de soluções?

ii. Que condição a matriz A deve satisfazer para se aplicar a fatoração Cholesky?

iii. Para a matriz $B = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 5 & 7 \\ 3 & 7 & 14 \end{bmatrix}$, sabemos que ela admite uma fatoração LU, onde $L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix}$. Obtenha U e calcule a solução do sistema $Ax = b$ para $b_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ e $b_2 = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}$ através da decomposição LU.

Exercício 4.13 Modele o sistema de treliças descrito na Figura 4.33, de forma a obter as matrizes A, B, e C tal que:

$$\begin{bmatrix} \Delta l_1 \\ \Delta l_2 \\ \Delta l_3 \end{bmatrix} = A \begin{bmatrix} \Delta S_x \\ \Delta S_y \end{bmatrix} \quad \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} = B \begin{bmatrix} \Delta l_1 \\ \Delta l_2 \\ \Delta l_3 \end{bmatrix} \quad \begin{bmatrix} f_x \\ f_y \end{bmatrix} = C \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix}.$$

Obtenha portanto a matriz $M = CBA$ de maneira que:

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = M \begin{bmatrix} \Delta S_x \\ \Delta S_y \end{bmatrix}.$$

Para uma carga $f = [\ f_x \quad f_y\ ]^T = [\ 1N \quad 1.5N\ ]^T$ e constantes $k_1 = 30N/m$, $k_2 = 28N/m$ e $k_3 = 35N/m$, obtenha o deslocamento ΔS resolvendo o sistema de equações lineares utilizando a fatoração LU. (Sugestão: utilize Octave/Matlab para calcular a decomposição LU de M).

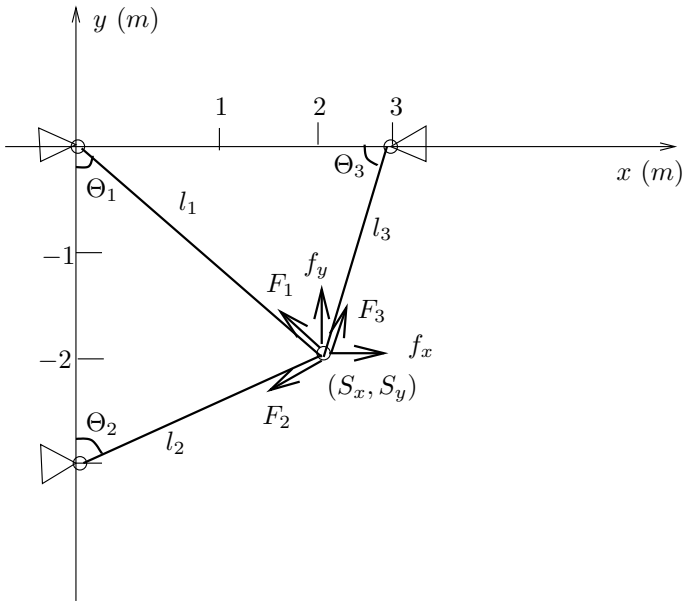


Figura 4.33: Sistema de treliças. O desenho não está em escala. $(S_x, S_y) = (2, -2)$.

Exercício 4.14 Questões de Verdadeiro/Falso.

i. Seja $\| \bullet \|_p$ uma norma vetorial definida por $p \in \mathbb{N}$. Então obrigatoriamente $\|x\|_\infty \geqslant \|x\|_p$ para qualquer $x \in \mathbb{R}^n$.

ii. Seja $A \in \mathbb{R}^{n\times n}$ uma matriz quadrada. Se A tem inversa, então $\|A^{-1}\| \geqslant \frac{1}{\|A\|}$.

iii. Seja $A \in \mathbb{R}^{n\times n}$ uma matriz quadrada que possui inversa. Suponha que $A = U\Lambda U^T$ tal que (i) $U^TU = I$ e $\|U\| = 1$ e (ii) Λ é uma matriz diagonal. Então obrigatoriamente $\kappa(A) = \kappa(\Lambda)$.

iv. Seja $A \in \mathbb{R}^{n\times n}$ uma matriz tal que $A = A^T$ e A^{-1} existe. Então, com certeza $A^k = A.A\ldots A$ (A multiplicada k vezes) é tal que $\kappa(A^k) = \kappa(A)^k$.

v. Seja $Ax = b$ um sistema de equações lineares onde $A \in \mathbb{R}^{n\times n}$ tal que $rank(A) = n$. Considere o método de Jacobi dado por $x^{(k+1)} = D^{-1}[(D - A)x^{(k)} + b]$, onde D é a matriz diagonal de A. Se $\|A\|_\infty < 1$ então o método de Jacobi é convergente para a solução.

vi. Seja $Ax = b$ um sistema de equações lineares onde $A \in \mathbb{R}^{m\times n}$, onde $m > n$, mas tal que $b \in range(A)$. Então o método de mínimos quadrados ($\min \|Ax - b\|$) pode ser aplicado e encontra uma solução x^* de erro mínimo, ou seja, $\|Ax^* - b\| = 0$.

vii. Seja $Ax = b$ um sistema de equações lineares onde $A \in \mathbb{R}^{m\times n}$ e tal que $n > m$ (há mais variáveis do que equações). Então o sistema tem uma ou mais soluções.

viii. Seja $Ax = b$ um sistema de equações lineares onde $A \in \mathbb{R}^{m\times n}$. Se $rank(A^T) = n$ e $b \in range(A)$ então o sistema tem solução única.

ix. Para $x \in \mathbb{R}^n$, seja $\|x\|_* = \min\{|x_j| : j = 1, \ldots, n\}$. Então $\|x\|_*$ é uma norma vetorial.

Exercício 4.15 Seja $A \in \mathbb{R}^{n\times n}$ uma matriz quadrada. Dados os autovalores $\lambda_1, \lambda_2, \ldots, \lambda_n$ de A, formule o problema de encontrar os autovetores como um problema (ou problemas) que pode ser resolvido numericamente. Como você resolveria este problema numérico?

Exercício 4.16 Seja $Ax = b$ um sistema de equações lineares, onde $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^{n\times 1}$ e $x \in \mathbb{R}^{n\times 1}$. Seja a matriz $D = diag(a_{11}, a_{22}, \ldots, a_{nn})$ (matriz com a diagonal de A) inversível, o que nos leva a obter o processo iterativo de Jacobi (clássico)

$$x^{(k+1)} = D^{-1}[b - (A - D)x^{(k)}]. \tag{4.56}$$

O processo iterativo de Jacobi obtido com o preparador unitário é dado por

$$x^{(k+1)} = b + (I - A)x^{(k)}. \tag{4.57}$$

Professor Kuhn afirma que se o processo iterativo de Jacobi (4.56) converge, a partir de um ponto inicial $x^{(0)} \in \mathbb{R}^n$, então o processo iterativo de Jacobi (4.57) também converge a partir de $x^{(0)}$.

Capítulo 5

Sistemas de Equações Não-Lineares, Otimização e Mínimos Quadrados

Neste capítulo estudaremos o problema de encontrar uma solução para um sistemas de equações não-lineares, para o qual vamos estender o algoritmo de Newton apresentado em capítulo anterior. Aplicações deste problema também serão discutidas. Noções básicas de otimização, incluindo conceitos de otimalidade e um algoritmo inspirado no método de Newton com retrocesso será proposto para resolver problemas de minimização. Trataremos ainda de problemas de interpolação, mínimos quadrados e minimização de norma. Este dois últimos problemas estão relacionados com os problemas de encontrar uma solução de um sistema de equações lineares sem solução (mínimos quadrados) e com infinitas soluções (minimização de norma).

5.1 Sistemas de Equações Não-Lineares

O problema de encontrar uma solução para um sistema de n equações não-lineares em n variáveis é expresso por:

$$\left\{\begin{array}{rcl} f_1(x_1, x_2, \ldots, x_n) & = & 0 \\ f_2(x_1, x_2, \ldots, x_n) & = & 0 \\ & \vdots & \\ f_n(x_1, x_2, \ldots, x_n) & = & 0 \end{array}\right. \quad \Rightarrow f(x) = 0 \tag{5.1}$$

onde $x \in \mathbb{R}^n$ e $f : \mathbb{R}^n \to \mathbb{R}^n$.

5.1.1 Matriz de Derivadas

Para $f : \mathbb{R}^n \to \mathbb{R}^n$ diferenciável, podemos calcular o gradiente de f, denotado por ∇f:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{5.2}$$

O gradiente ∇f é conhecido também por Jacobiano.

Exemplo: Cálculo do Jacobiano

Considere a função $f : \mathbb{R}^2 \to \mathbb{R}^2$ definida por:

$$f(x) = \begin{bmatrix} e^{2x_1+x_2} - x_1 \\ x_1^2 - x_2 \end{bmatrix}$$

O gradiente de f obtido segundo a expressão (5.2) é:

$$\nabla f(x) = \begin{bmatrix} 2e^{2x_1+x_2} - 1 & e^{2x_1+x_2} \\ 2x_1 & -1 \end{bmatrix}$$

5.1.2 Linearização

Podemos desenvolver a expansão de Taylor para uma função $f : \mathbb{R}^n \to \mathbb{R}^n$, da mesma forma que desenvolvemos para uma função $f : \mathbb{R} \to \mathbb{R}$. Assumindo que $f = [f_1, \ldots, f_n]^T$, vamos primeiramente obter a expansão de Taylor de primeira ordem para f_i em torno de um ponto $\bar{x}$:

$$\begin{aligned} f_i(x) &\approx f_i(\bar{x}) + \frac{\partial f_i}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f_i}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2) + \ldots + \frac{\partial f_i}{\partial x_n}(\bar{x})(x_n - \bar{x}_n) \\ &= f_i(\bar{x}) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(\bar{x})(x_j - \bar{x}_j) \\ &= f_i(\bar{x}) + \nabla f_i(\bar{x})^T (x - \bar{x}) \end{aligned}$$

Portanto

$$\begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix} \approx \begin{bmatrix} f_1(\bar{x}) \\ f_2(\bar{x}) \\ \vdots \\ f_n(\bar{x}) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f_1}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2) + \ldots + \frac{\partial f_1}{\partial x_n}(\bar{x})(x_n - \bar{x}_n) \\ \frac{\partial f_2}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f_2}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2) + \ldots + \frac{\partial f_2}{\partial x_n}(\bar{x})(x_n - \bar{x}_n) \\ \vdots \\ \frac{\partial f_n}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f_n}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2) + \ldots + \frac{\partial f_n}{\partial x_n}(\bar{x})(x_n - \bar{x}_n) \end{bmatrix}$$

$$\approx \begin{bmatrix} f_1(\bar{x}) \\ f_2(\bar{x}) \\ \vdots \\ f_n(\bar{x}) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \frac{\partial f_1}{\partial x_2}(\bar{x}) & \ldots & \frac{\partial f_1}{\partial x_n}(\bar{x}) \\ \frac{\partial f_2}{\partial x_1}(\bar{x}) & \frac{\partial f_2}{\partial x_2}(\bar{x}) & \ldots & \frac{\partial f_2}{\partial x_n}(\bar{x}) \\ \vdots & & & \\ \frac{\partial f_n}{\partial x_1}(\bar{x}) & \frac{\partial f_n}{\partial x_2}(\bar{x}) & \ldots & \frac{\partial f_n}{\partial x_n}(\bar{x}) \end{bmatrix} \begin{bmatrix} (x_1 - \bar{x}_1) \\ (x_2 - \bar{x}_2) \\ \vdots \\ (x_n - \bar{x}_n) \end{bmatrix}$$

De forma mais compacta,

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x})(x - \bar{x}) \\ &= f(\bar{x}) + \nabla f(\bar{x})\Delta x \end{aligned}$$

Exemplo: Expansão de Taylor

Linearizando a função do exemplo anterior em torno do ponto $x = [0\ 0]$ obtemos:

$$f(x) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

5.1.3 Aplicação: Circuito Estático Não-Linear

Aqui vamos considerar o problema de analisar o comportamento em regime permanente do circuito não-linear ilustrado na Figura 5.1. Dois resistores não-lineares caracterizados pelas equações:

$$\begin{aligned} i_1 &= g(v_1) \\ i_2 &= g(v_2) \end{aligned}$$

onde $g(v)$ é uma função não-linear da tensão através do resistor.

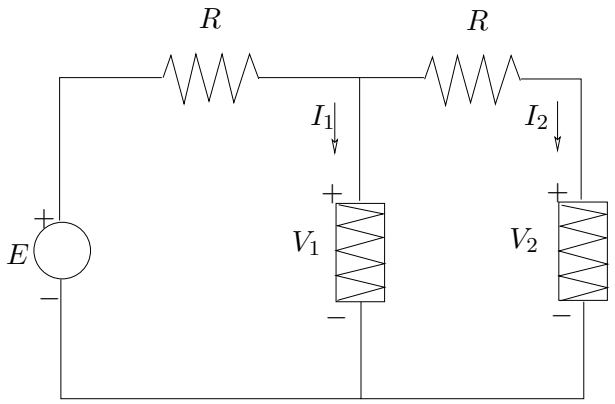


Figura 5.1: Circuito elétrico com componentes não-lineares

Desenvolvendo as equações associadas ao circuito obtemos:

$$\begin{cases} E &= R(i_1+i_2)+v_1 \\ E &= R(i_1+i_2)+Ri_2+v_2 \end{cases} \Rightarrow$$

$$\begin{cases} E &= Rg(v_1)+Rg(v_2)+v_1 \\ E &= Rg(v_1)+Rg(v_2)+Rg(v_2)+v_2 \end{cases} \Rightarrow$$

$$\begin{cases} \frac{E-v_1}{R} &= g(v_1)+g(v_2) \\ \frac{E-v_2}{R} &= g(v_1)+2g(v_2) \end{cases} \Rightarrow$$

$$\begin{cases} g(v_1)+g(v_2)+\frac{v_1-E}{R} &= 0 \\ g(v_1)+2g(v_2)+\frac{v_2-E}{R} &= 0 \end{cases}$$

Portanto, os valores de tensão através dos resistores podem ser obtidos resolvendo o sistema de equações não-lineares acima.

Outra maneira de se obter um sistema de equações equivalente segue

$$\begin{cases} v_{R_1}+v_1 &= E \Rightarrow v_{R_1}=E-v_1 \\ i_{R_1} &= \frac{v_{R_1}}{R}=\frac{E-v_1}{R} \end{cases}$$

$$\begin{cases} v_{R_2}+v_2 &= v_1 \Rightarrow v_{R_2}=v_1-v_2 \\ i_{R_2} &= \frac{v_{R_2}}{R}=\frac{v_1-v_2}{R} \end{cases}$$

$$\begin{cases} i_{R_2} = i_2 &\Rightarrow & i_2 &= \frac{v_1-v_2}{R} \\ &\Rightarrow & g(v_2) &= \frac{v_1-v_2}{R} \\ &\Rightarrow & g(v_2)-\frac{v_1-v_2}{R} &= 0 \end{cases}$$

$$\begin{cases} i_{R_1} &= i_1+i_2 \\ \frac{E-v_1}{R} &= g(v_1)+\frac{v_1-v_2}{R} \end{cases} \Rightarrow \quad g(v_1)+\frac{v_1-E}{R}+\frac{v_1-v_2}{R} = 0$$

Portanto, o problema se reduz a encontrar uma raiz para o sistema de equações abaixo:

$$\begin{aligned} g(v_1) + \frac{v_1 - E}{R} + \frac{v_1 - v_2}{R} &= 0 \\ g(v_2) - \frac{v_1 - v_2}{R} &= 0 \end{aligned}$$

Podemos então colocar o sistema (5.3) na forma:

$$\left\{ \begin{array}{rcl} f_1(v_1, v_2) &=& g(v_1) + \frac{v_1 - E}{R} + \frac{v_1 - v_2}{R} = 0 \\ f_2(v_1, v_2) &=& g(v_2) - \frac{v_1 - v_2}{R} = 0 \end{array} \right.$$

ou, de forma mais compacta,

$$f(v) = \left[\begin{array}{c} f_1(v_1, v_2) \\ f_2(v_1, v_2) \end{array} \right] = 0$$

Vamos agora calcular o Jacobiano de f em torno de um ponto $\hat{v} = (\hat{v}_1, \hat{v}_2)$:

$$\nabla f(\hat{v}_1, \hat{v}_2) = \left[\begin{array}{cc} g'(\hat{v}_1) + 2/R & -1/R \\ -1/R & g'(\hat{v}_2) + 1/R \end{array} \right]$$

Portanto, a expansão de Taylor de primeira ordem para f em torno do ponto $(\hat{v}_1, \hat{v}_2)$ tem a forma:

$$f(v_1, v_2) \approx f(\hat{v}_1, \hat{v}_2) + \nabla f(\hat{v}_1, \hat{v}_2). \left[\begin{array}{c} v_1 - \hat{v}_1 \\ v_2 - \hat{v}_2 \end{array} \right]$$

Com base na expansão de Taylor acima, podemos desenvolver o método de Newton para o problema de análise do circuito não-linear. Desejamos que:

$$f(\hat{v}_1, \hat{v}_2) + \nabla f(\hat{v}_1, \hat{v}_2). \left[\begin{array}{c} v_1 - \hat{v}_1 \\ v_2 - \hat{v}_2 \end{array} \right] = 0 \Rightarrow \left[\begin{array}{c} v_1 \\ v_2 \end{array} \right] = \left[\begin{array}{c} \hat{v}_1 \\ \hat{v}_2 \end{array} \right] - \nabla f^{-1}(\hat{v}_1, \hat{v}_2). f(\hat{v}_1, \hat{v}_2)$$

assumindo que $\nabla f(\hat{v}_1, \hat{v}_2)$ é inversível.

5.1.4 Algoritmo de Newton

Newton$(f, x^{(0)}, \epsilon)$

1: $k \leftarrow 0$
2: **while** $\|f(x^{(k)})\| > \epsilon$ **do**
3: $\quad x^{(k+1)} \leftarrow x^{(k)} - \nabla f^{-1}(x^{(k)}).f(x^{(k)})$
4: $\quad k \leftarrow k + 1$
5: **end while**
6: Return $x^{(k)}$

Observe que cada iteração requer uma avaliação de $f(x)$ (i.e., n funções escalares) e de $\nabla f(x)$ (i.e., n^2 derivadas parciais). Note também que o método de Newton assume que $\nabla f(x^k)$ é não singular pois, caso contrário, o método não está definido. Na prática, calcula-se $x^k - \nabla f^{-1}(x^k).f(x^k)$ através da solução de um sistema de equações lineares:

a) Calcule $J_k = \nabla f(x^k)$ e $g_k = f(x^k)$

b) Resolva o sistema:

$$\begin{aligned} x^{k+1} = x^k - J_k^{-1} g_k \quad &\Rightarrow \quad J_k(x^{k+1} - x^k) = -g_k \\ &\Rightarrow \quad J_k \Delta x_k = -g_k, \ \Delta x_k = x^{k+1} - x^k \end{aligned}$$

c) Obtenha o próximo iterando:

$$x^{k+1} = x^k + \Delta x_k$$

Exemplo 1

Considere o problema de encontrar uma solução para o sistema de equações não-lineares abaixo:

$$\left\{ \begin{array}{rcccc} f_1(x_1, x_2) & = & \log(x_1^2 + 2x_2^2 + 1) - 0.5 & = & 0 \\ f_2(x_1, x_2) & = & x_2 - x_1^2 + 0.2 & = & 0 \end{array} \right.$$

O sistema possui duas equações e duas variáveis, conforme ilustração na Figura 5.2. As soluções para o sistema são $(0.70, 0.29)$ e $(-0.70, 0.29)$.

Exemplo 2: Rastreamento com Radares

Aqui vamos considerar o cenário onde dois radares estão localizados em posições conhecidas, (p_1, q_1) e (p_2, q_2), os quais podem determinar a distância de suas posições até uma aeronave que está se deslocando dentro do espaço aéreo. Denotando por ρ_1 e ρ_2 as distâncias destes radares até a aeronave, o problema é determinar a localização (x, y) da aeronave. O problema é ilustrado na Figura 5.3. Podemos colocar o problema na forma de duas equações e duas variáveis:

$$\left\{ \begin{array}{rcccc} f_1(x, y) & = & \sqrt{(p_1 - x)^2 + (q_1 - y)^2} - \rho_1 & = & 0 \\ f_2(x, y) & = & \sqrt{(p_2 - x)^2 + (q_2 - y)^2} - \rho_2 & = & 0 \end{array} \right.$$

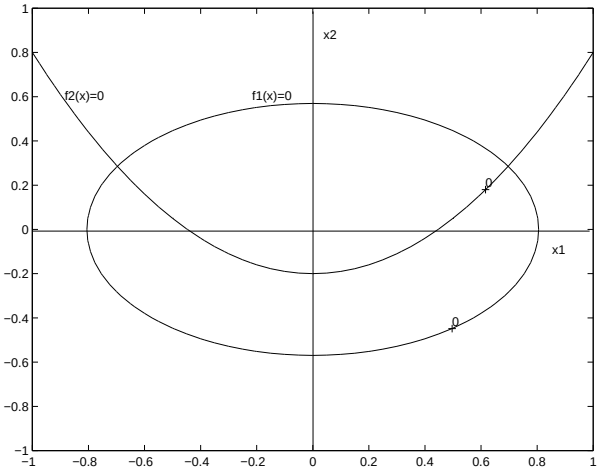


Figura 5.2: Ilustração das curvas de nível de duas funções não lineares

Assumindo que $(x, y) \neq (p_1, q_1)$ e $(x, y) \neq (p_2, q_2)$, o Jacobiano de $f = (f_1, f_2)$ pode ser obtido como segue:

$$\nabla f(x,y) = \begin{bmatrix} \partial f_1/\partial x & \partial f_1/\partial y \\ \partial f_2/\partial x & \partial f_2/\partial y \end{bmatrix} = \begin{bmatrix} \frac{x-p_1}{\sqrt{(p_1-x)^2+(q_1-y)^2}} & \frac{y-q_1}{\sqrt{(p_1-x)^2+(q_1-y)^2}} \\ \frac{x-p_2}{\sqrt{(p_2-x)^2+(q_2-y)^2}} & \frac{y-q_2}{\sqrt{(p_2-x)^2+(q_2-y)^2}} \end{bmatrix}$$

Fazendo

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \text{ e } f(z) = \begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix}$$

O método de Newton assume a forma:

$$z^{k+1} = z^k - \nabla f^{-1}(z^k).f(z^k)$$

5.1.5 Convergência do Método de Newton

Teorema 5.1 *Se $\nabla f(x)$ é não-singular e x^0 é suficientemente próximo de uma solução x^*, $f(x^*) = 0$, então existe uma constante $\epsilon > 0$ tal que:*

$$\|x^{k+1} - x^*\| \leqslant \epsilon \|x^k - x^*\|^2$$

Em outras palavras, se as condições do teorema são satisfeitas, o algoritmo de Newton converge com taxa quadrática. Contudo, na prática, não sabemos o valor de ϵ e não sabemos quão próximo de x^* o iterando inicial x^0 deve estar.

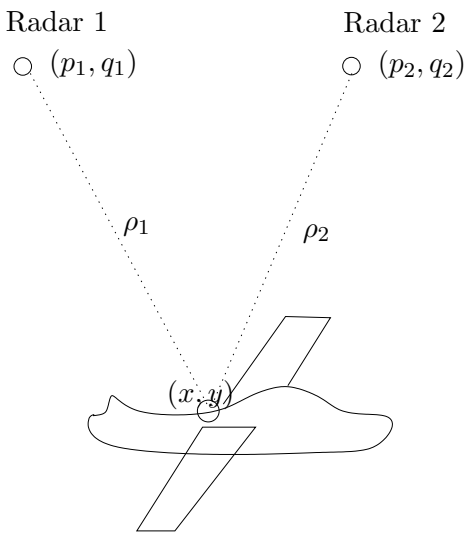


Figura 5.3: Ilustração de dois radares que detectam a distância até uma aeronave

5.2 Minimização Irrestrita

Um problema geral com aplicações diversas é o problema de encontrar um mínimo de uma função qualquer sem restrições. Em notação matemática, o problema toma a forma:

$$\begin{array}{l} \text{Minimize} \quad f(x) \\ x \in \mathbb{R}^n \end{array}$$

onde $f : \mathbb{R}^n \to \mathbb{R}$ é uma função e $x \in \mathbb{R}^n$ é um vetor de variáveis de decisão.

Definição 5.1 *(Mínimo Global) x^* é um ótimo global se $f(x^*) \leqslant f(x)$ para todo $x \in \mathbb{R}^n$.*

Definição 5.2 *(Mínimo Local) x^* é otimo local se existe $\epsilon > 0$, tal que $f(x^*) \leqslant f(x)$ para todo $x \in \mathbb{R}^n$ tal que $\|x - x^*\| \leqslant \epsilon$.*

A Figura 5.4 ilustra os conceitos de mínimos locais e globais. Os pontos x_1, x_2, e x_4 são mínimos locais enquanto que o ponto x_3 é um mínimo global.

Definição 5.3 *(Valor Ótimo) O valor ótimo (ou mínimo) de f é o maior $\alpha \in \mathbb{R}$ tal que $f(x) \geqslant \alpha$ para todo $x \in \mathbb{R}^n$, neste caso $\alpha = \min f(x)$.*

- *Se x^* é uma solução ótima, então $f(x^*) = \min f(x)$.*
- *Se $f(x)$ não tem limite inferior, então dizemos que $\min f(x) = -\infty$.*

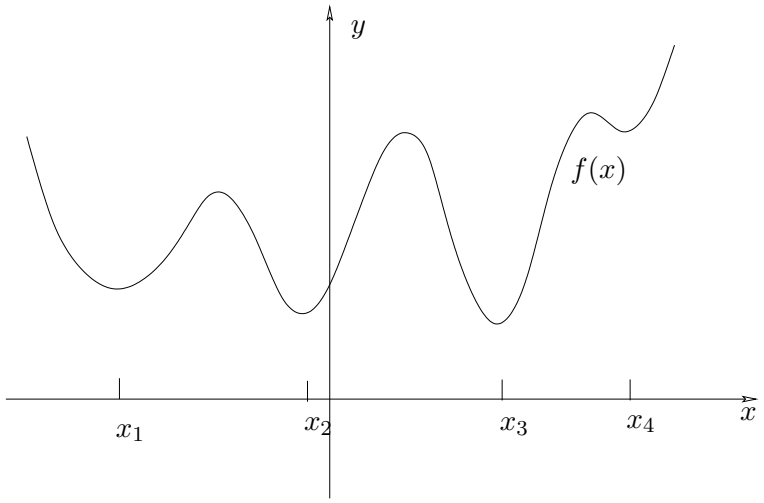


Figura 5.4: Ilustração de mínimos locais e globais de uma função $f(x)$

5.2.1 Exemplos

Os exemplos de funções abaixo ilustram os conceitos de valor ótimo e solução ótima.

1) $f(x) = (x-1)^2$ tem valor ótimo obtido com $x^* = 1$.

2) $f(x) = \frac{1}{x^2+1}$ tem valor ótimo 0, mas não é atingido por nenhuma solução finita.

3) $f(x) = x$ não tem valor ótimo, $\min f(x) = -\infty$.

5.2.2 Condições de Otimalidade para Problemas com uma Variável

Seja $f(x)$ uma função de uma variável, $f : \mathbb{R} \to \mathbb{R}$, para a qual desejamos encontrar um mínimo. Vamos assumir que f é duas vezes diferenciável.

Proposição 5.1 *(Condição necessária) Se x^* é um ótimo local, então $f'(x^*) = 0$ e $f''(x^*) \geqslant 0$.*

Prova: Primeiro, vamos mostrar que $f'(x^*) = 0$. Suponha que existe mínimo local x^* tal que $f'(x^*) \neq 0$. Pela expansão de Taylor de primeira ordem, temos que:

$$f(x) = f(x^*) + f'(x^*)(x - x^*)$$

para $|x - x^*|$ suficientemente pequeno. Defina $x = -f'(x^*)\delta + x^*$ com $\delta > 0$ suficientemente pequeno para que a expansão seja válida. Então:

$$\begin{aligned} f(x) &= f(x^*) + f'(x^*)(-f'(x^*)\delta + x^* - x^*) \\ &= f(x^*) - f'(x^*)^2\delta \\ &< f(x^*) \qquad\qquad [\text{pois } f'(x^*) \neq 0 \text{ e } \delta > 0] \end{aligned}$$

Isto nos leva a concluir que x^* não é mínimo local, contradizendo a hipótese. Concluímos que se x^* é mínimo local então obrigatoriamente $f'(x^*) = 0$.

Segundo, vamos mostrar que $f''(x^*) \geqslant 0$. Suponha que x^* é mínimo local e que $f''(x^*) < 0$. Pela expansão de Taylor de segunda orgem, temos que:

$$\begin{aligned} f(x) &= f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2}f''(x^*)(x - x^*)^2 \\ &= f(x^*) + \frac{1}{2}f''(x^*)(x - x^*)^2 \qquad [\text{pois } f'(x^*) = 0] \end{aligned}$$

para $|x - x^*|$ suficientemente pequeno. Defina $x = \delta + x^*$ com $|\delta| > 0$ suficientemente pequeno para garantir a validade da expansão de Taylor. Então:

$$\begin{aligned} f(x) &= f(x^*) + \frac{1}{2}f''(x^*)(x - x^*)^2 \\ &= f(x^*) + \frac{1}{2}f''(x^*)(\delta + x^* - x^*)^2 \\ &= f(x^*) + \frac{1}{2}f''(x^*)\delta^2 \\ &< f(x^*) \qquad\qquad [\text{pois } f''(x^*) < 0 \text{ e } \delta^2 > 0] \end{aligned}$$

implicando em existir x na vizinhança de x^* com objetivo estritamente menor. Mas isto significa que x^* não pode ser um mínimo local, contradizendo a hipótese. Portanto, se x^* é um mínimo local então obrigatoriamente $f''(x^*) \geqslant 0$. ■.

Proposição 5.2 *(Condição suficiente) Se x^* satisfaz $f'(x^*) = 0$ e $f''(x^*) > 0$, então x^* é um ótimo local.*

Prova: Seja x^* um ponto que satisfaz as condições $f'(x^*) = 0$ e $f''(x^*) > 0$. Seja $x = \delta + x^*$ um ponto na vizinhança de x^* com $\delta \neq 0$. Com $|\delta|$

suficientemente pequeno, pela expansão de Taylor de segunda ordem, vale:

$$\begin{aligned} f(x) &= f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2}f''(x^*)(x - x^*)^2 \\ &= f(x^*) + \frac{1}{2}f''(x^*)(\delta + x^* - x^*)^2 \\ &= f(x^*) + \frac{1}{2}f''(x^*)\delta^2 \\ &> f(x^*) \qquad\qquad [\text{pois } f''(x^*) > 0 \text{ e } \delta^2 > 0] \end{aligned}$$

Logo, existe uma vizinhança $N(x^*)$ em torno de x^* tal que $f(x^*) < f(x)$ para todo $x \in N(x^*)\backslash\{x^*\}$, satisfazendo as condições de otimalidade local. ■

Definição 5.4 *(Função convexa) Uma função $f : \mathbb{R} \to \mathbb{R}$ é convexa se para todo $x, y \in \mathbb{R}$ e $\alpha \in [0, 1]$ valer:*

$$f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y)$$

Proposição 5.3 *Se f é convexa, então todo o mínimo local é também um mínimo global.*

Prova: Seja x^* um mínimo local e suponha, por absurdo, que x^* não seja um mínimo global. Seja $\tilde{x}$ um mínimo global e considere o segmento de reta que une os pontos x^* e $\tilde{x}$, ou seja,

$$x = \alpha\tilde{x} + (1 - \alpha)x^* \text{ onde } \alpha \in (0, 1].$$

Por convexidade de f, temos que:

$$\begin{aligned} f(x) &\leqslant \alpha f(\tilde{x}) + (1 - \alpha)f(x^*) \\ &< \alpha f(x^*) + (1 - \alpha)f(x^*) \qquad [\text{Pois } f(\tilde{x}) < f(x^*)] \\ &= f(x^*) \end{aligned}$$

Qualquer vizinhança $\mathcal{N}$ de x^* contém um pedaço do segmento de reta acima definido e, portanto, sempre existirá um x em $\mathcal{N}$ que satisfaça a desigualdade $f(x) < f(x^*)$. Conclui-se que x^* não é um ótimo local, contradizendo a hipótese. ■

Proposição 5.4 *Uma função diferenciável f é convexa se e somente se $f(z) \geqslant f(x) + f'(x)(z - x)$ para todo $x, z \in \mathbb{R}$.*

Prova: (Necessidade) Suponha que f é convexa. Uma vez que f é diferenciável, vale:

$$\lim_{\alpha\to 0}\frac{f(x+\alpha(z-x))-f(x)}{\alpha} = f'(x)(z-x) \tag{5.3}$$

Sendo f convexa, temos que:

$$\begin{aligned} & f(x+\alpha(z-x)) \leqslant \alpha f(z)+(1-\alpha)f(x), && \forall\alpha\in[0,1] \\ \Rightarrow\ & f(x+\alpha(z-x))-f(x) \leqslant \alpha(f(z)-f(x)), && \forall\alpha\in[0,1] \\ \Rightarrow\ & \frac{f(x+\alpha(z-x))-f(x)}{\alpha} \leqslant f(z)-f(x), && \forall\alpha\in[0,1] \end{aligned} \tag{5.4}$$

Fazendo $\alpha\to 0$ em (5.4) e usando (5.3), deduzimos que:

$$\begin{aligned} f(z)-f(x) &\geqslant \lim_{\alpha\to 0}\frac{f(x+\alpha(z-x))-f(x)}{\alpha} \\ &= f'(x)(z-x) \\ \Rightarrow f(z) &\geqslant f(x)+f'(x)(z-x) \end{aligned}$$

(Suficiência) Suponha que $f(z)\geqslant f(x)+f'(x)(z-x)$ para todo $x,z\in\mathbb{R}$. Seja $z=\alpha x+(1-\alpha)y$ para $x,y\in\mathbb{R}$ e $\alpha\in[0,1]$. A partir desta desigualdade, deduzimos:

$$\begin{aligned} f(x) &\geqslant f(z)+f'(z)(x-z) \\ f(y) &\geqslant f(z)+f'(z)(y-z) \end{aligned}$$

Multiplicando a primeira por α, multiplicando a segunda por $(1-\alpha)$ e somando os resultados, obtemos:

$$\begin{aligned} \alpha f(x)+(1-\alpha)f(y) &\geqslant f(z)+(\alpha(x-z)+(1-\alpha)(y-z))f'(z) \\ &= f(z)+(\alpha x+(1-\alpha)y-\alpha z-(1-\alpha)z)f'(z) \\ &= f(z)+(z-z)f'(z) \\ &= f(z) \\ &= f(\alpha x+(1-\alpha)y) \end{aligned}$$

demonstrando que f é convexa. ■

A Proposição 5.4 é ilustrada na Fig. 5.5. Para qualquer ponto x, note que a reta $y(z)=f(x)+f'(x)(z-x)$ define um limite inferior para $f(z)$, ou seja, $y(z)\leqslant f(z)$.

Proposição 5.5 *Seja f uma função duas vezes continuamente diferenciável. Se $f''(x)\geqslant 0$ para todo x, então f é convexa.*

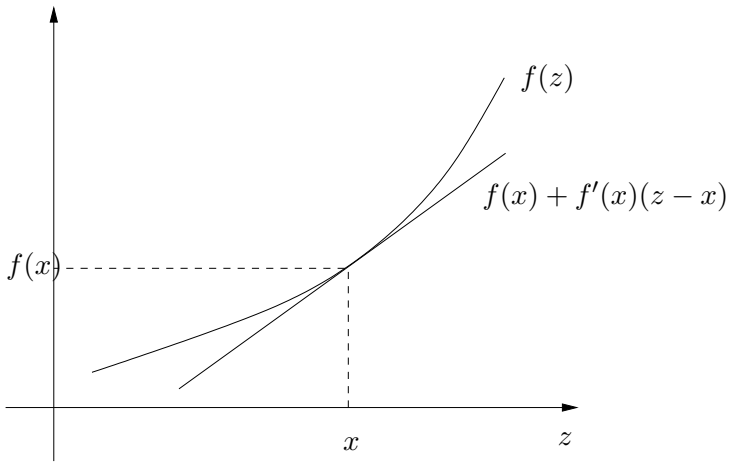


Figura 5.5: Ilustração de função convexa: $f(z) \geqslant f(x) + f'(x)(z - x)$ para todo $x, z \in \mathbb{R}$.

Prova: A partir da expansão de Taylor de segunda ordem, para quaisquer x e z tem-se:

$$f(z) = f(x) + f'(x)(z - x) + \frac{1}{2} f''(x + \alpha(z - x))(z - x)^2$$

para alguma $\alpha \in [0, 1]$. Uma vez que $f''(y) \geqslant 0$ para todo y, deduzimos que:

$$f(z) \geqslant f(x) + f'(x)(z - x)$$

Logo, segue da Proposição 5.4 que f é convexa. ■

Exemplos de Funções Convexas

1) $f(x) = ax^2 + bx + c$ com $a > 0$, pois $f'(x) = 2ax + b$ e $f''(x) = 2a$. Portanto, $f''(x) > 0$ para todo o lugar. $x^* = \frac{-b}{2a}$ é um ótimo local único.

2) Para $f(x) = \log(e^x + e^{-x})$ temos que

$$\begin{array}{rcl} f'(x) & = & \frac{e^x - e^{-x}}{e^x + e^{-x}} \\ f''(x) & = & \frac{4}{(e^x + e^{-x})^2} \end{array}$$

   Portanto, $f''(x) > 0$ para todo x e $x^* = 0$ é um ótimo global.

3) Para $f(x) = x^4$, temos

$$\begin{array}{rcl} f'(x) & = & 4x^3 \\ f''(x) & = & 12x^2 \end{array}$$

   Logo $f''(x) \geqslant 0$ em todo o lugar e $x^* = 0$ é um otimo global único.

4) Para $f(x) = x^3$, podemos verificar que $f''(x) = 6x$. Embora $f'(0) = 0$ e $f''(0) = 0$, $x = 0$ não é um ótimo local.

5.3 Minimização de Funções de uma Variável: Método de Newton

Considere o problema de minimização irrestrita em uma variável

$$\begin{array}{l} \text{Minimize} \quad f(x) \\ x \in \mathbb{R} \end{array}$$

onde $f : \mathbb{R} \to \mathbb{R}$. Podemos realizar a busca por um ponto x que satisfaz as condições de otimalidade de primeira ordem aplicando o método de Newton na equação $f'(x) = 0$. Neste caso, o algoritmo de Newton consistirá dos passos dados abaixo.

Newton-Minimize$(f, x^{(0)}, \epsilon)$
1: $k \leftarrow 0$
2: **while** $|f'(x^{(k)})| > \epsilon$ **do**
3: $\quad x^{(k+1)} \leftarrow x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}$
4: $\quad k \leftarrow k + 1$
5: **end while**
6: return $x^{(k)}$

Alguns problemas podem surgir na aplicação da estratégia acima:

- o método converge somente se iniciarmos com um ponto x_0 próximo da solução x^*, onde $f'(x^*) = 0$; e
- pode convergir para um ponto de máximo ou de inflexão.

5.3.1 Interpretação de uma Iteração

Considere a linearização $f'(x')$ em torno do ponto x, conforme desenvolvimento abaixo:

$$f'(x') \approx f_{lin}(x') = f'(x) + f''(x)(x' - x)$$

Fazendo $f_{lin}(x') = 0$ teremos

$$\begin{aligned} f_{lin}(x') = 0 \quad &\Leftrightarrow \quad f'(x) + f''(x)(x' - x) = 0 \\ &\Leftrightarrow \quad x' = x - \frac{f'(x)}{f''(x)} \end{aligned}$$

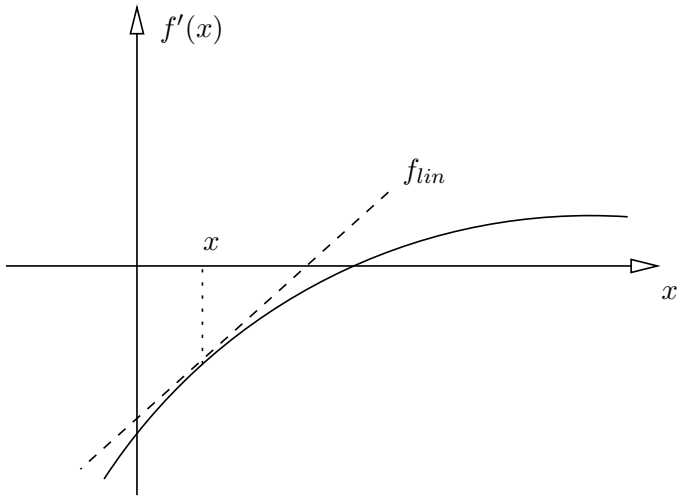


Figura 5.6: Visão geométrica do método de Newton aplicado ao problema $f'(x) = 0$

5.3.2 Uma Segunda Interpretação

Tomemos a aproximação quadrática $f(x')$ de $f(x)$:

$$f(x') \approx h(x') = f(x) + f'(x)(x' - x) + \frac{1}{2}f''(x)(x' - x)^2$$

Faça x^+ o mínimo de $h(x')$, ou seja,

$$\begin{aligned} h'(x') = 0 \quad &\Rightarrow \quad f'(x) + f''(x)(x' - x) = 0 \\ &\Rightarrow \quad x' = x - \frac{f'(x)}{f''(x)} \end{aligned}$$

Exemplo

Tomando como exemplo a função $f(x) = \log(e^x + e^{-x})$, calculamos as derivadas de primeira e segunda ordem:

$$\begin{aligned} f'(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} \\ f''(x) &= \frac{4}{(e^x + e^{-x})^2} \end{aligned}$$

As funções $f(x)$ e $f'(x)$ podem ser observadas na Figura 5.7.

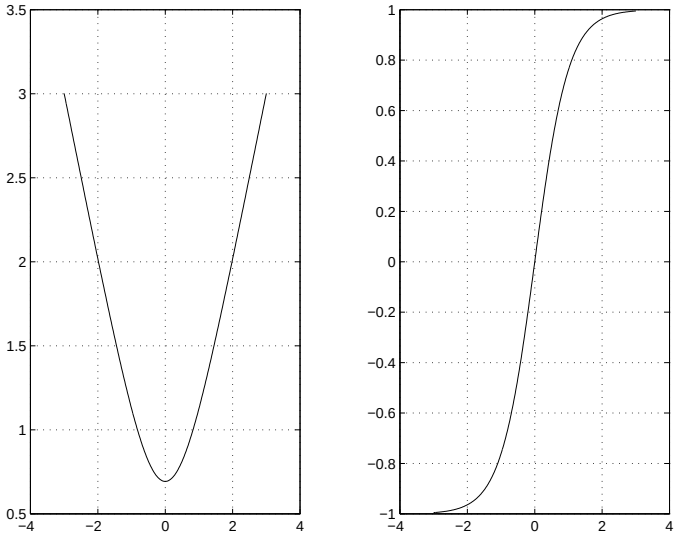


Figura 5.7: Ilustração gráfica das funções $f(x)$ e $f'(x)$

5.4 Método de Newton com Retrocesso ("Backtracking") para Funções Convexas

Convex-Backtracking-Minimize$(f, x^{(0)}, \epsilon, \alpha)$

$k \leftarrow 0$
while $|f'(x^{(k)})| > \epsilon$ **do**
  $v_k \leftarrow -\frac{f'(x_k)}{f''(x_k)}$
  while $f(x_k + v_k) > f(x_k) + \alpha f'(x_k) v_k$ **do**
    $v_k \leftarrow \frac{v_k}{2}$
  end while
  $x_{k+1} \leftarrow x_k + v_k$
  $k \leftarrow k + 1$
end while
return $x^{(k)}$

O parâmetro $\alpha \in (0, 0.5)$ e $\epsilon > 0$. O laço mais interno é dito "*backtracking*": decresça o passo v_k até que $f(x_k + v_k) \leqslant f(x_k) + \alpha f'(x_k) v_k$

A Figura 5.8 ilustra o funcionamento do método. O passo v_k que induz o próximo iterando $x_{k+1} = x_k + v_k$ só é aceito se o decréscimo induzido na função é maior do que o previsto pela reta $f(x_k) + \alpha f'(x_k) v_k$. Note que

para v_k suficientemente pequeno, a condição imposta é satisfeita. Portanto, o método começa com o passo de Newton $v_k = -f'(x_k)/f''(x_k)$ e reduz o passo até que $v_k \leqslant v_{max}$.

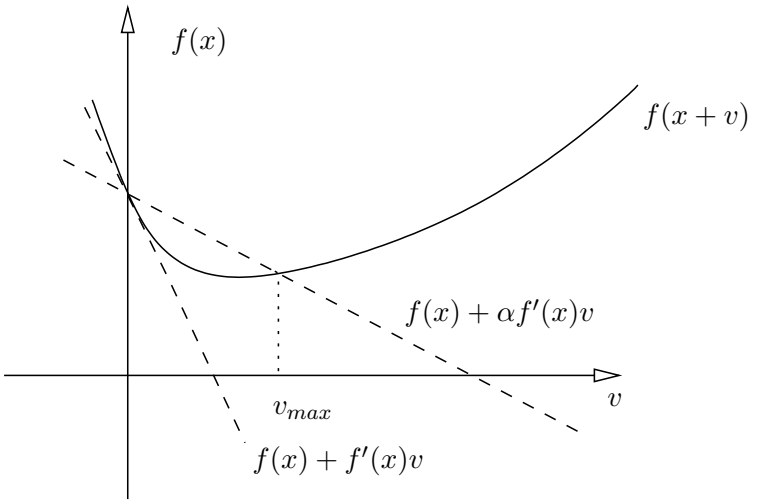


Figura 5.8: Interpretação geométrica do método de Newton para minimização de funções convexas

5.4.1 Convergência do Método de Newton com Retrocesso

Teorema 5.2 *Se $f''(x) > 0$ em todo o lugar e f tem um ponto ótimo global x^*, então o método de Newton com* backtracking *converge globalmente para x^* com taxa de convergência quadrática.*

5.5 Algoritmo de Newton para Minimização de Funções Não Convexas

Seja $x^+ = x - \frac{f'(x)}{f''(x)}$ o minimizador da aproximação de segunda ordem $f_2(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(x)(y-x)^2$ da função $f(x)$ em torno do ponto x. Para o exemplo ilustrado na Figura 5.9, a solução x^+ é pior do que a solução corrente x. Na verdade, a aplicação pura do método vair convergir para um ponto de máximo. Uma maneira de contornar este problema consiste em seguir a lógica abaixo:

- Se $f''(x) \leqslant 0$, faça $x^+ = x - f'(x)$ (*tome o passo dado pelo método de descenso*)

- Se $f''(x) > 0$, faça $x^+ = x - \frac{f'(x)}{f''(x)}$ (*aceita o passo de Newton*)

A estratégia esboçada acima pode ser sintetizada no algoritmo abaixo.

Nonconvex-Backtracking-Minimize$(f, x^{(0)}, \epsilon, \alpha)$

$k \leftarrow 0$
while $|f'(x^{(k)})| > \epsilon$ **do**
  if $f''(x_k) > 0$ **then**
    $v_k \leftarrow -\frac{f'(x_k)}{f''(x_k)}$
  else
    $v_k \leftarrow -f'(x_k)$
  end if
  while $f(x_k + v_k) > f(x_k) + \alpha f'(x_k) v_k$ **do**
    $v_k \leftarrow \frac{v_k}{2}$
  end while
  $x_{k+1} \leftarrow x_k + v_k$
  $k \leftarrow k + 1$
end while
return $x^{(k)}$

Observações:

a) os iterandos satisfazem $f(x_{k+1}) < f(x_k)$

b) o algoritmo converge para um mínimo local

c) próximo do mínimo local x^*, se $f''(x^*) > 0$, o algoritmo toma passos de Newton que garante convergência quadrática local.

5.6 Mínimos Quadrados e Ajuste de Curvas

Consideremos inicialmente um sistema com mais equações do que variáveis:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ a_{31}x_1 + a_{32}x_2 + \ldots + a_{3n}x_n = b_3 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{cases}$$

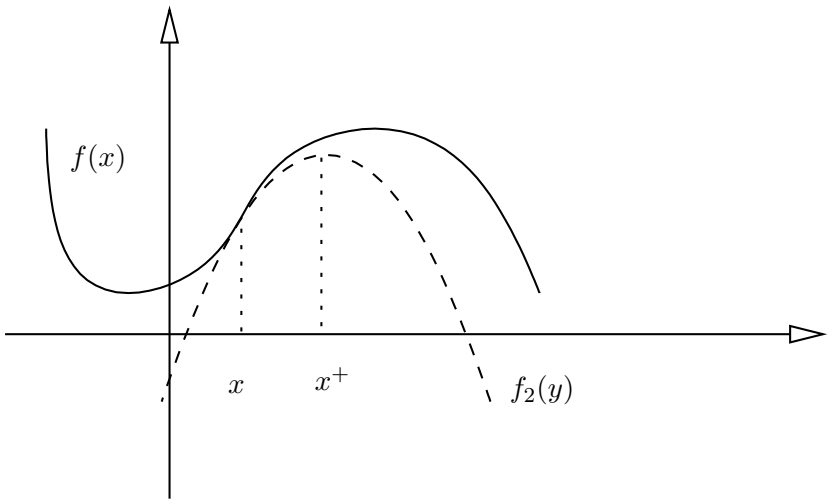


Figura 5.9: Problemas do método de Newton aplicado à minimização de funções não convexas.

o qual pode ser colocado na forma mais compacta como:

$$Ax = b$$

onde $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^{n \times 1}$, e $b \in \mathbb{R}^{m \times 1}$.

Vamos assumir que não existe solução x tal que $Ax = b$.

5.6.1 Solução por Mínimos Quadrados

O problema é encontrar uma solução aproximada de $Ax = b$, formalmente, queremos minimizar o erro que pode ser colocado em notação matemática como segue:

$$P: \quad \underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad \|Ax - b\|$$

onde $r = Ax - b$ é chamado de resíduo. A solução x para P com menor erro residual $\|r\| = \|Ax - b\|$ é chamada de solução de mínimos quadrados. O problema P pode ser colocado em uma forma equivalente P':

$$P': \quad \underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad \|Ax - b\|^2 = (Ax - b)^T(Ax - b) = \sum_{i=1}^{m}(a_i^T x - b_i)^2$$

onde a_i^T é a i-ésima linha de A.

Exemplo

Considere o problema de encontrar uma solução para um sistema de três equações a duas variáveis:

$$\begin{cases} 2x_1 & & = & 1 \\ -x_1 & + \quad x_2 & = & 0 \\ & 2x_2 & = & -1 \end{cases} \tag{5.5}$$

A solução por mínimos quadrados para (5.5) pode ser colocada na forma abaixo:

$$\underset{x_1, x_2}{\text{Minimize}} \quad f = (2x_1 - 1)^2 + (x_2 - x_1)^2 + (2x_2 + 1)^2$$

Para encontrar a solução ótima, as derivadas parciais de f devem ser nulas, ou seja, o gradiente ∇f de f deve ser nulo. Portanto:

$$\nabla f(x) = 0 \quad \Rightarrow \quad \begin{aligned} \frac{\partial f}{\partial x_1} &= 10x_1 - 2x_2 - 4 = 0 \\ \frac{\partial f}{\partial x_2} &= -2x_1 + 10x_2 + 4 = 0 \end{aligned}$$

Logo, a solução é:

$$x_1 = \frac{1}{3} \text{ e } x_2 = -\frac{1}{3}$$

5.6.2 Ajuste de Curvas

O problema geral de ajuste de curvas pode ser colocado como segue. Ajuste a curva dada pela função:

$$g(t) = x_1 g_1(t) + x_2 g_2(t) + \ldots + x_n g_n(t)$$

aos dados $(t_1, y_1), \ldots, (t_m, y_m)$, ou seja, gostaríamos que:

$$\begin{aligned} g(t_1) &= y_1 \\ g(t_2) &= y_2 \\ \vdots &= \vdots \\ g(t_m) &= y_m \end{aligned}$$

onde $g_i(t) : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$, são funções quaisquer mas conhecidas; tais funções são ditas bases. Note que:

a) as variáveis do problema são $x_1, x_2, \ldots, x_n$; e

b) tipicamente $m \gg n$.

Algumas aplicações de ajuste de curvas são:

a) suavização de dados;

b) desenvolvimento de modelos para dados amostrais; e

c) extrapolação.

5.6.3 Ajuste de Curvas: Um Problema de Mínimos Quadrados

Podemos colocar o problema de ajuste de curvas como um problema de mínimos quadrados, o qual assume a forma:

$$\begin{aligned} \text{Minimize} \quad f &= \sum_{i=1}^{m} (g(t_i) - y_i)^2 \\ &= \sum_{i=1}^{m} [x_1 g_1(t_i) + x_2 g_2(t_i) + \ldots + x_n g_n(t_i) - y_i]^2 \end{aligned}$$

que, por sua vez, pode ser expresso de forma matricial:

$$\begin{array}{c} \text{Minimize} \quad \|Ax - b\|^2 \\ x \in \mathbb{R}^n \end{array}$$

onde

$$A = \begin{bmatrix} g_1(t_1) & g_2(t_1) & \ldots & g_n(t_1) \\ g_1(t_2) & g_2(t_2) & \ldots & g_n(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ g_1(t_m) & g_2(t_m) & \ldots & g_n(t_m) \end{bmatrix} \text{ e } b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

5.6.4 Exemplo: Ajuste de Polinômios

O problema consiste em ajustar o polinômio $g(t) = x_1 + x_2 t + x_3 t^2 + \ldots + x_n t^{n-1}$ aos dados $(t_1, y_1), \ldots, (t_m, y_m)$, onde $m \gg n$, sendo que desejamos que

$$\begin{aligned} g(t_1) &= y_1 \\ g(t_2) &= y_2 \\ \vdots &= \vdots \\ g(t_m) &= y_m \end{aligned}$$

Em forma matricial, desejamos resolver o problema:

$$\begin{array}{c} \text{Minimize} \quad \|Ax - b\|^2 \\ x \in \mathbb{R}^n \end{array}$$

onde

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \dots & t_2^{n-1} \\ 1 & t_3 & t_3^2 & \dots & t_3^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^{n-1} \end{bmatrix} \text{ e } b = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}$$

Se $m = n$ e o sistema tiver solução única, pode-se utilizar um método de solução de equações lineares.

5.6.5 Exemplo: Ajuste de Curva

Aqui desejamos encontrar um polinômio que melhor aproxima a função $f(t) = \frac{1}{1+25t^2}$ no intervalo $[-1, 1]$. Não sabemos de antemão quantos pontos deverão ser utilizados para realizar a aproximação, ou seja, não sabemos o valor de m. Além disso, não sabemos qual grau deve ser adotado para o polinômio, portanto vamos ter que realizar alguns experimentos numéricos e traçar gráficos que nos ajudem a identificar os valores mais adequados para m (número de pontos amostrais) e n (sendo $n-1$ o grau do polinômio interpolador).

A Figura 5.10 ilustra a aproximação de $f(t)$ obtida ao ajustarmos $p(t) = x_1 + x_2t + x_3t^2 + \ldots + x_nt^{n-1}$ a um conjunto de $m = 5$ pontos sendo o grau do polinômio $n - 1 = 4$ ($n = 5$).

A Figura 5.11 ilustra a aproximação de $f(t)$ com um polinômio de grau $n - 1 = 14$ ($n = 15$) ao conjunto de $m = 17$ pontos.

A Figura 5.12 ilustra a aproximação de $f(t)$ com um polinômio de grau $n - 1 = 4$ ($n = 5$) a um conjunto de $m = 65$ pontos.

Por fim, a Figura 5.13 mostra graficamente a aproximação de $f(t)$ obtida com um polinômio $p(t)$ de grau $n-1 = 14$ ($n = 15$) a um conjunto de $m = 65$ pontos.

5.6.6 Identificação de Sistemas

O problema de identificação de sistema trata da busca de modelos de sistemas que não conhecemos seus elementos e conexões. Temos apenas valores de entrada e seus respectivos valores de saída. Para o sistema ilustrado na Figura 5.14, por exemplo, podemos realizar um conjunto de experimentos

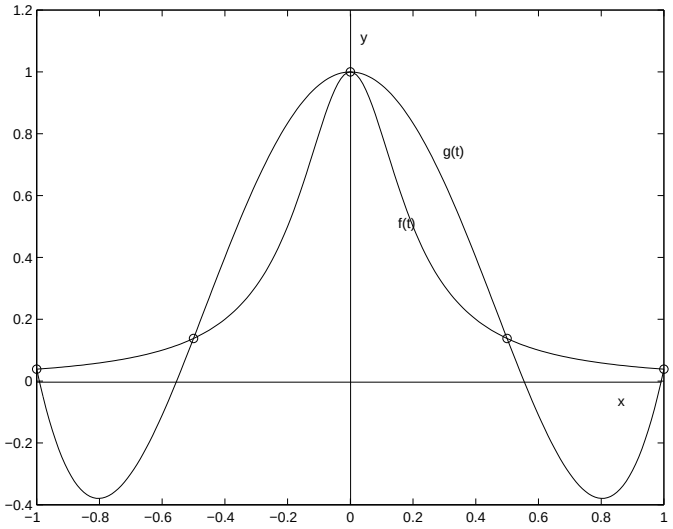


Figura 5.10: $n = 5$ e $m = 5$

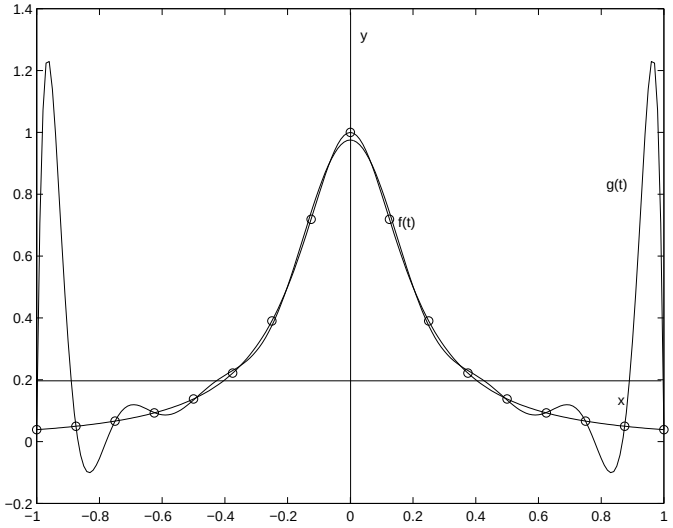


Figura 5.11: $n = 15$ e $m = 17$

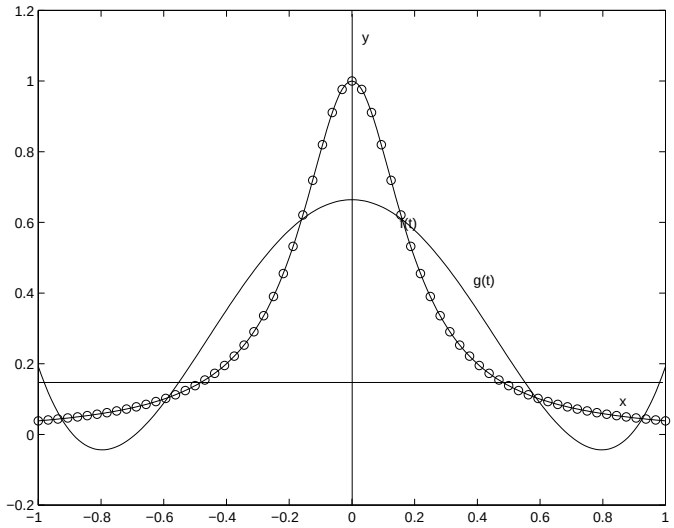


Figura 5.12: $n = 5$ e $m = 65$

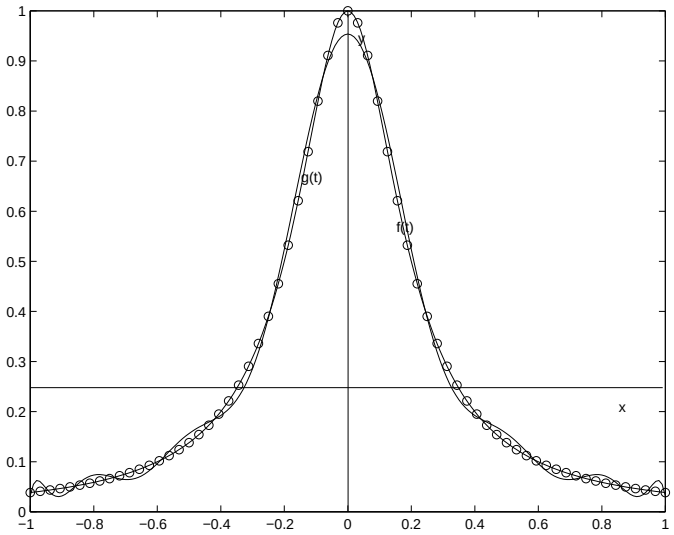


Figura 5.13: $n = 15$ e $m = 65$

que consistem em injetar valores $u(t), t = 0, \ldots, N$, e observar as respectivas saídas temporais $y(t), t = 0, \ldots, N$.

Figura 5.14: Exemplo de sistema caixa-preta

Considere o exemplo de entradas e saídas de um sistema qualquer conforme Figura 5.15. Um modelo simples e amplamente utilizado para modelagem de sistema é dado por:

$$\hat{y}(t) = h_0 u(t) + h_1 u(t-1) + h_2 u(t-2) + \cdots + h_n u(t-n) \tag{5.6}$$

Este modelo é conhecido por *moving average* com n atrasos. Assim $\hat{y}(t)$ é o valor de predição de $y(t)$ produzido pelo modelo para a entrada corrente, $u(t)$, e as n entradas anteriores, $u(t-1)$, $u(t-2)$, $\ldots$, $u(t-n)$. Portanto, a predição é uma combinação linear da entrada corrente e das n entradas anteriores, sendo $h_0, h_1, \ldots, h_n$ os parâmetros que definem a combinação linear das entradas.

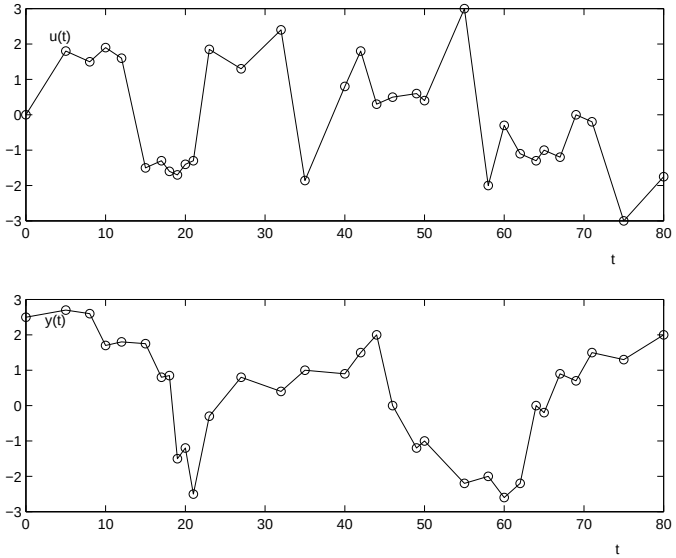


Figura 5.15: Exemplo de entradas e saídas de um sistema

5.6.7 Estimação por Meio de Mínimos Quadrados

Aqui vamos considerar o problema de encontrar um modelo (i.e., parâmetros $h_0, \ldots, h_n$) tal que o erro de predição seja minimizado:

$$\begin{array}{ll} \text{Minimize} & E = \left[\sum_{t=n}^{t=N} (\hat{y}(t) - y(t))^2\right]^{\frac{1}{2}} \\ h_0, \ldots, h_n & \end{array} \tag{5.7}$$

onde: $\hat{y}(t) = h_0 u(t) + h_1 u(t-1) + h_2 u(t-2) + \ldots + h_n u(t-n)$. Note que o erro E pode ser expresso na forma matricial, fazendo:

$$\begin{aligned} E &= \left[\sum_{t=n}^{t=N} (\hat{y}(t) - y(t))^2\right]^{\frac{1}{2}} \\ &= \|Ax - b\| \end{aligned}$$

$$A = \begin{bmatrix} u(n) & u(n-1) & u(n-2) & \ldots & u(0) \\ u(n+1) & u(n) & u(n-1) & \ldots & u(1) \\ u(n+2) & u(n+1) & u(n) & \ldots & u(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u(N) & u(N-1) & u(N-2) & \ldots & u(N-n) \end{bmatrix}$$

$$x = \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} \text{ e } b = \begin{bmatrix} y_n \\ y_{n+1} \\ y_{n+2} \\ \vdots \\ y_N \end{bmatrix}$$

Logo, o problema (5.7) pode ser colocado como um problema de mínimos quadrados:

$$\begin{array}{ll} \text{Minimize} & \|Ax - b\|^2 \\ x \in \mathbb{R}^{n+1} & \end{array}$$

5.6.8 Identificação do Sistema Motor Taco-Gerador

O motor taco-gerador é um sensor de velocidade largamente utilizado na indústria. Ele consiste em um gerador de tensão ligado a um motor por uma correia ou engrenagem como indica a Fig. 5.16. Conforme varia a velocidade angular do motor varia a tensão gerada pelo taco-gerador. Em motores elétricos, a velocidade angular do motor depende da tensão aplicada sobre ele. Desta forma, pode-se estabelecer uma relação causal entre a entrada de

tensão u aplicada ao motor e a tensão de saída y mostrada pelo taco-gerador (função da velocidade angular ω do motor).

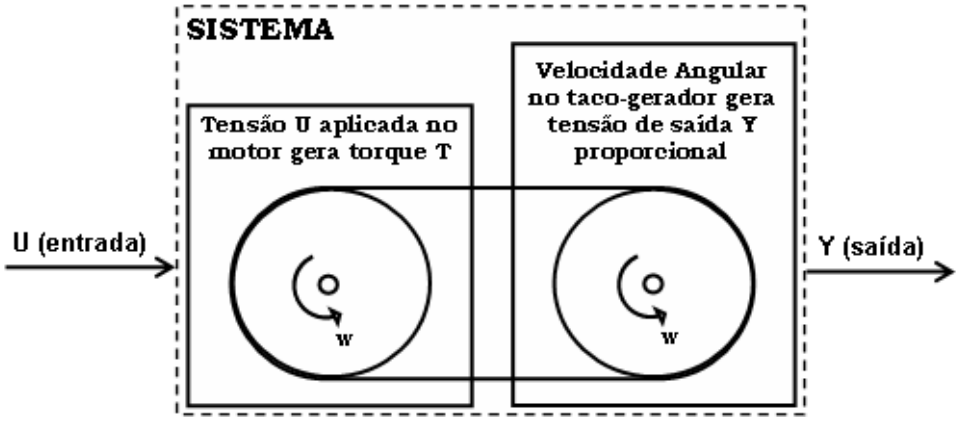


Figura 5.16: Motor taco-gerador.

A relação entre a tensão de entrada u e a tensão de saída y pode ser obtida de duas formas diferentes:

Modelagem Físico-Matemática: baseada nas leis da física e da matemática que caracterizam o sistema. No caso do motor taco-gerador, aplicando as Leis de Newton, chegaríamos à seguinte equação diferencial:

$$J \cdot L \frac{d^2y(t)}{dt^2} + (J \cdot R + B \cdot L) \frac{dy(t)}{dt} + (B \cdot R + K_a \cdot K_b)\, y(t) = \frac{K_a}{K_c}\, u(t)$$

onde:

- J é o momento de inércia do motor;
- K_a é a constante de proporcionalidade entre a corrente na armadura e o torque no motor;
- K_b é a constante de proporcionalidade entre a velocidade angular e a força contra-eletromotriz do motor elétrico;
- K_c é a constante de proporcionalidade entre a velocidade angular do gerador e a tensão y;
- B é o atrito do motor elétrico;
- L é a indutância da armadura do motor elétrico;
- R é a resistência da armadura do motor elétrico

A modelagem físico-matemática de sistemas será objeto de estudo no Capítulo 7 que trata da síntese e solução numérica de equações diferenciais.

Modelagem Experimental ou Identificação de Sistemas: onde a partir da entrada e da saída do sistema estima-se um modelo matemático similar ao modelo real. Na Fig. 5.17 são dadas as curvas da entrada $u(t)$ e da saída $y(t)$ obtidas através de um ensaio em laboratório.

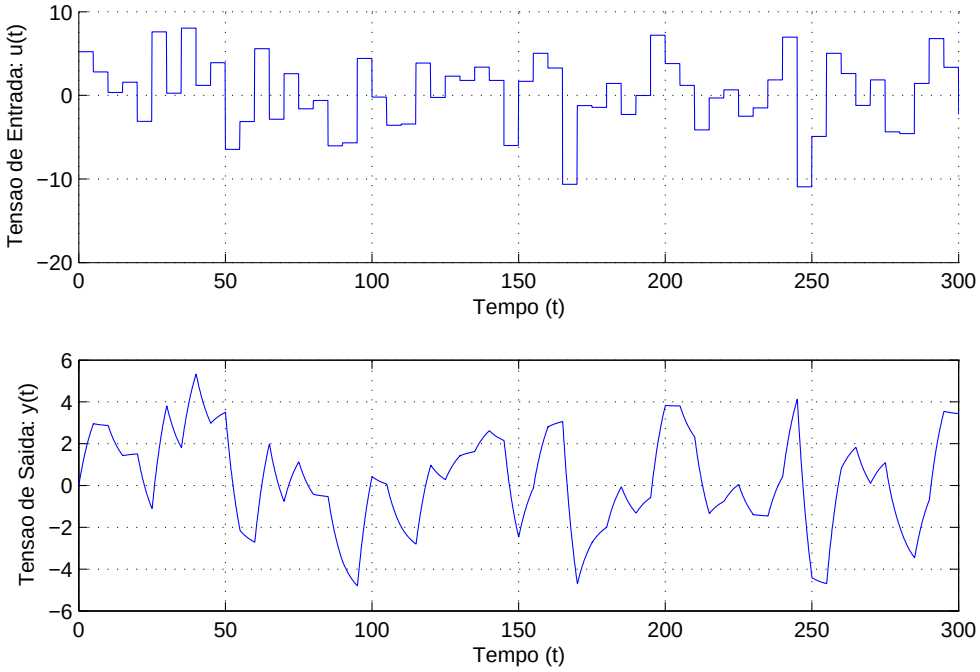


Figura 5.17: Motor taco-gerador.

O modelo de predição $\hat{y}(t)$ está ilustrado nas Figs. 5.18 e 5.19 para $n = 3$ e $n = 10$, respectivamente. Como pode ser observado nestas figuras, o modelo de predição não é satisfatório como indicam as normas dos vetores de erro, $e = y(t) - \hat{y}(t)$, que apresentam o valor $\|e\| = 118.5214$ para o caso $n = 3$ e $\|e\| = 107.4303$ para o caso $n = 10$.

O modelo de predição pode ser aprimorado utilizando-se a equação alternativa a seguir:

$$\begin{aligned}\hat{y}(t) = h_0u(t) + h_1u(t-1) + h_2u(t-2) + \cdots + h_nu(t-n) + \\ + w_1y(t-1) + w_2y(t-2) + \cdots + w_ny(t-n)\end{aligned} \quad (5.8)$$

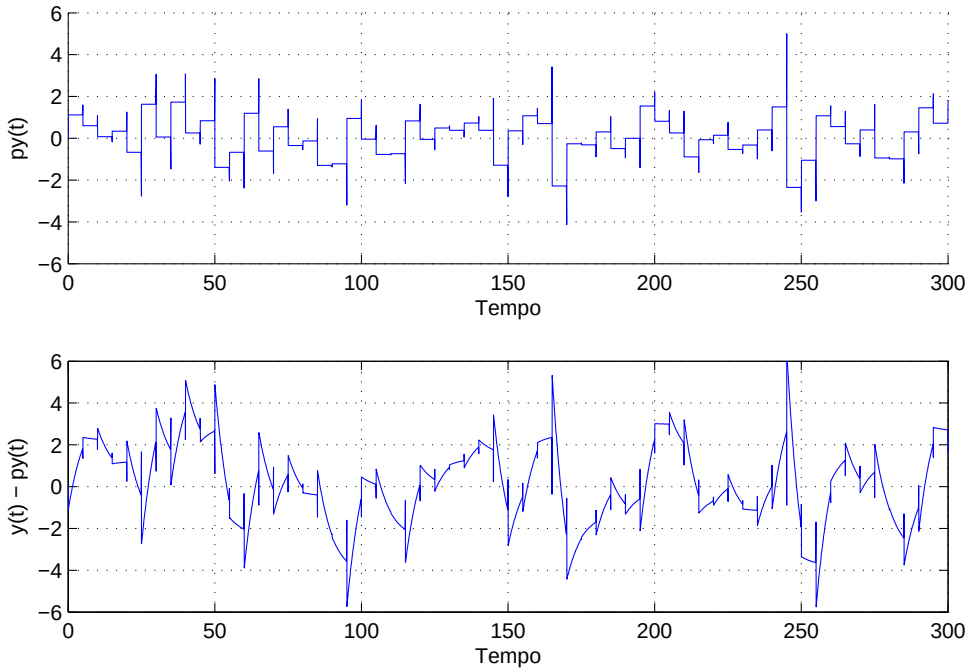


Figura 5.18: Modelo de predição $\hat{y}(t)$ obtido com a equação (5.6) e $n = 3$. O vetor de erros $e = y(t) - \hat{y}(t)$ tem norma $\|e\| = 118.5214$.

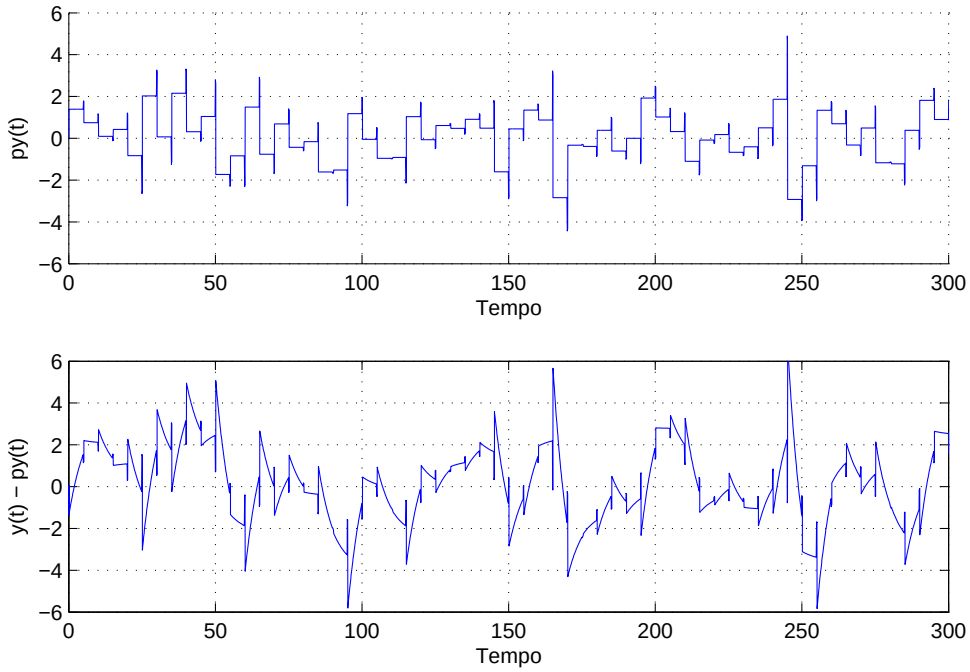


Figura 5.19: Modelo de predição $\hat{y}(t)$ obtido com a equação (5.6) e $n = 10$. O vetor de erros $e = y(t) - \hat{y}(t)$ tem norma $\|e\| = 107.4303$.

que leva em consideração as saídas passadas além das entradas passadas. Os resultados obtidos com a equação de predição alternativa (5.8) estão ilustrados nas Figs. 5.20 e 5.21. Podemos observar uma qualidade satisfatória do modelo alternativo.

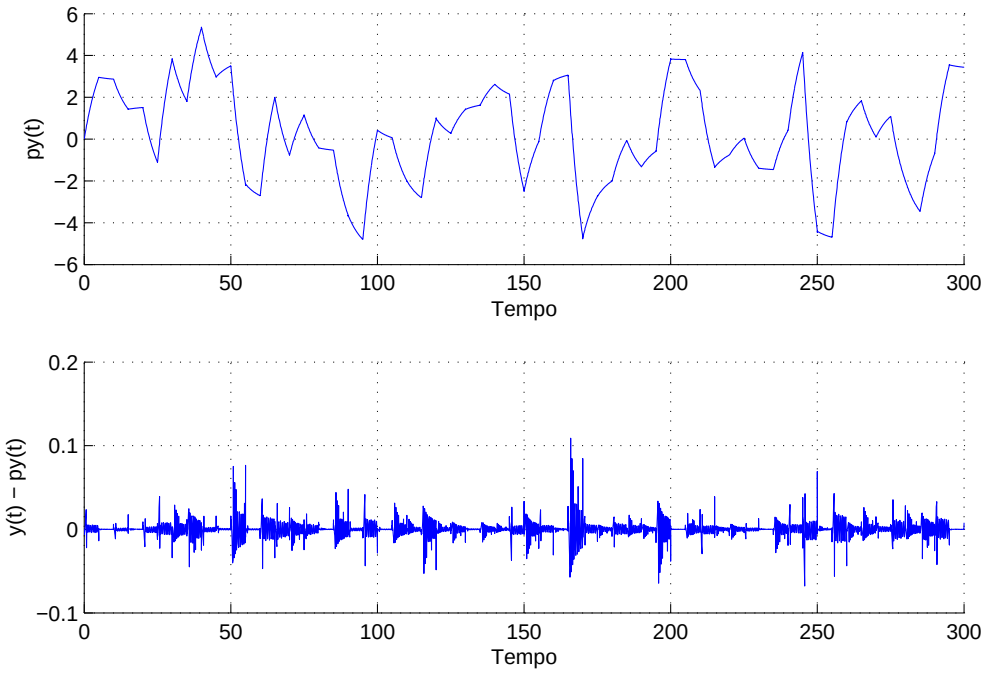


Figura 5.20: Modelo de predição $\hat{y}(t)$ obtido com a equação (5.8) e $n = 3$. O vetor de erros $e = y(t) - \hat{y}(t)$ tem norma $\|e\| = 0.5495$.

5.6.9 Resolução de Problemas de Mínimos Quadrados

Nesta seção vamos desenvolver uma fórmula com a solução explícita do problema de mínimos quadrados. Primeiramente, vamos lembrar que o problema tem a forma:

$$\begin{array}{l} \text{Minimize} \quad \|Ax - b\|^2 \\ x \in \mathbb{R}^n \end{array}$$

Seja $A = [a_1, a_2, \ldots, a_n]$ onde $a_j \in \mathbb{R}^{m\times 1}$ é o vetor correspondente a j-ésima coluna de A. Então a distância $\|Ax - b\|$ é a distância do vetor b ao vetor

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

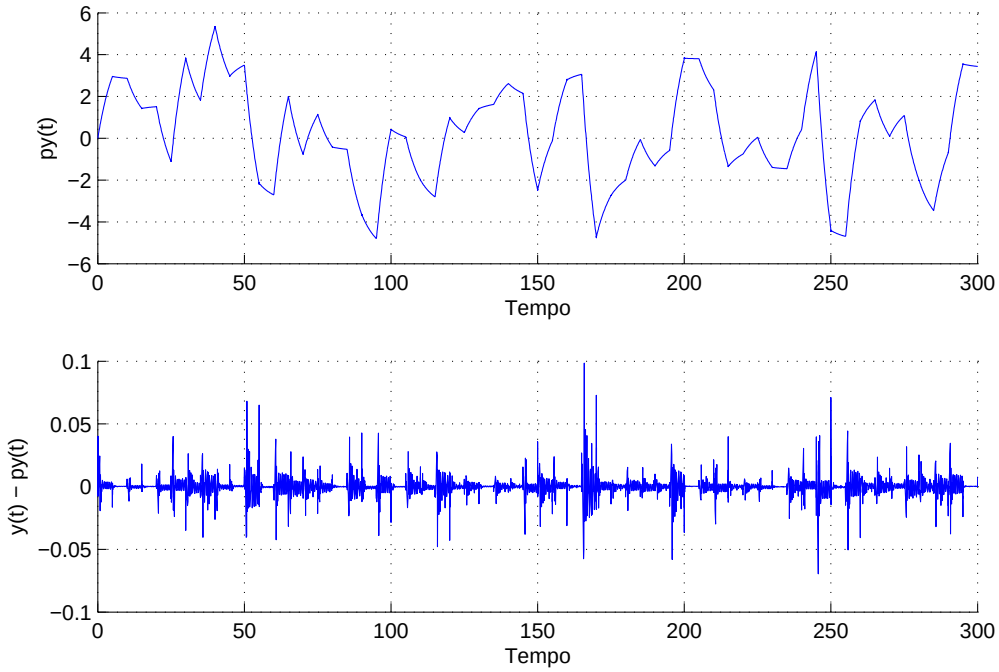


Figura 5.21: Modelo de predição $\hat{y}(t)$ obtido com a equação (5.8) e $n = 10$. O vetor de erros $e = y(t) - \hat{y}(t)$ tem norma $\|e\| = 0.4778$.

Logo o problema é encontrar uma combinação das colunas de A que seja mais próxima de b. A combinação Ax que minimiza o erro é precisamente a projetação de b no espaço gerado pelas colunas de A ($range(A) = \{Ax : x \in \mathbb{R}^n\}$). Abaixo ilustramos este aspecto geométrico do problema.

Exemplo

Considere o problema de mínimos quadrados onde:

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix} \text{ e } b = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$$

Graficamente, as colunas de A, o vetor b e a solução x_{LS} do problema de mínimos quadrados são ilustrados na Figura 5.22.

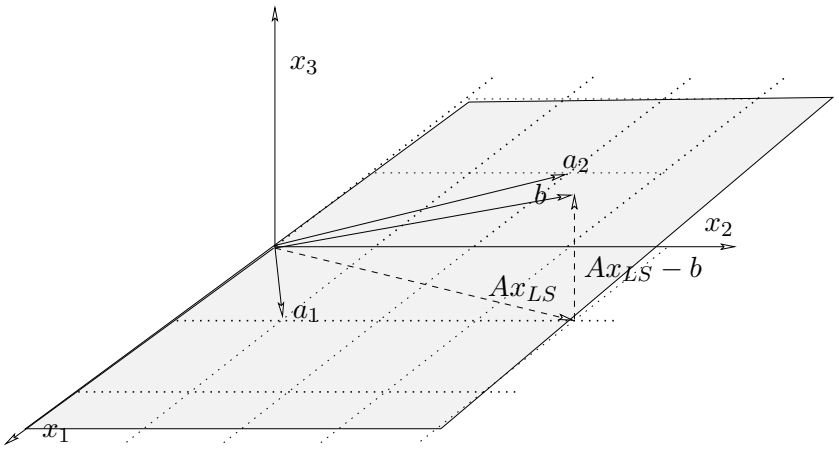


Figura 5.22: Ilustração geométrica do problema de mínimos quadrados

Solução do Problema de Mínimos Quadrados

Para o problema de mínimos quadrados dado na Seção 5.6.9, se $rank(A) = n$, então a solução é única e dada por:

$$x_{LS} = (A^T A)^{-1} A^T b$$

Em outras palavras, se $x \neq x_{LS}$ então $\|Ax - b\|^2 > \|Ax_{LS} - b\|^2$. Note que assumimos que $A \in \mathbb{R}^{m \times n}$ com $m \geqslant n$. Note que podemos encontrar a solução x_{LS} resolvendo o sistema de equações lineares em n variáveis e n equações dado por:

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

Primeiro, vamos mostrar que A^TA é não-singular.

$$\begin{aligned} A^TAx = 0 &\Rightarrow x^TA^TAx = 0 \\ &\Rightarrow \|Ax\|^2 = 0 \\ &\Rightarrow Ax = 0 \\ &\Rightarrow x = 0 \text{ pois } A \text{ tem posto completo} \end{aligned}$$

Portanto, a única solução para $A^TAx = 0$ é $x = 0$ que implica que A^TA não é singular.

Segundo, vamos mostrar que se $x \neq x_{LS}$ tem-se $\|Ax-b\|^2 > \|Ax_{LS}-b\|^2$.

$$\begin{aligned} \|Ax-b\|^2 &= \|Ax-b+Ax_{LS}-Ax_{LS}\|^2 \\ &= \|A(x-x_{LS})+(Ax_{LS}-b)\|^2 \\ &= \|A(x-x_{LS})\|^2+\|(Ax_{LS}-b)\|^2 \\ &\qquad \text{pois } A(x-x_{LS}) \text{ e } (Ax_{LS}-b) \text{ são ortogonais} \\ &> \|Ax_{LS}-b\|^2 \text{ já que } x \neq x_{LS} \end{aligned}$$

A ortogonalidade dos vetores acima pode ser verifica como segue:

$$\begin{aligned} (A(x-x_{LS}))^T(Ax_{LS}-b) &= (x-x_{LS})^TA^T(Ax_{LS}-b) \\ &= (x-x_{LS})^T(A^TAx_{LS}-A^Tb) \\ &= (x-x_{LS})^T(A^Tb-A^Tb) \\ &= 0 \end{aligned}$$

Interpretação Geométrica

A solução ótima para o problema de mínimos quadrados, x_{LS}, é tal que Ax_{LS} é a projeção de b sobre o espaço gerado pelas colunas de A, como ilustra a Figura 5.23

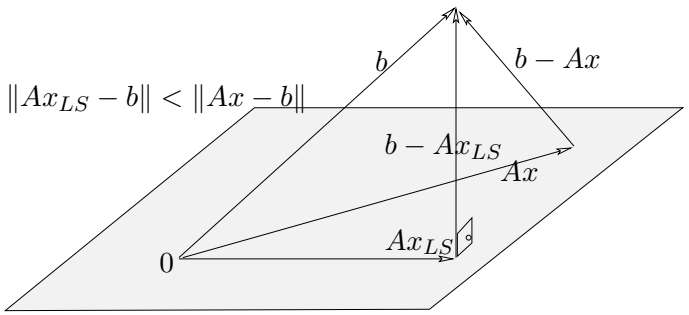


Figura 5.23: Ilustração geométrica do problema de mínimos quadrados

Note que $A^TAx_{LS} = A^Tb \rightarrow (Ax)^T(b-Ax_{LS}) = 0$. Portanto, $(b-Ax_{LS}) \perp Ax$, para todo $x \in \mathbb{R}^n$. Assim, $\|b-Ax\|^2 = \|b-Ax_{LS}\|^2 + \|A(x-x_{LS})\|$

Interpretação Alternativa

Defina $f : \mathbb{R} \to \mathbb{R}$ como:

$$f(x) = \|b - Ax\|^2 = \sum_{i=1}^{m} (b_i - a_i^T x)^2$$

onde:

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix}$$

sendo $a_i \in \mathbb{R}^{n \times 1}$ o vetor correspondente a i-ésima linha de A, $i = 1, \ldots, m$. Note que f é uma função diferenciável em n variáveis. A solução x_{LS} é um minimizador de f, caracterizado por:

$$\nabla f(x_{LS}) = \begin{bmatrix} \frac{\partial f(x_{LS})}{\partial x_1} \\ \frac{\partial f(x_{LS})}{\partial x_2} \\ \vdots \\ \frac{\partial f(x_{LS})}{\partial x_n} \end{bmatrix} = 0$$

Quais são as derivadas parciais de $f(x) = \sum_{i=1}^{m} (a_i^T x - b_i)^2$? Pode ser verificado que $\partial f / \partial x_j$ é dado por:

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{m} 2(a_i^T x - b_i) a_{ij}$$

Logo, o gradiente de f no ponto x é:

$$\begin{aligned}
\nabla f(x) &= \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \\
&= \sum_{i=1}^{m} 2(a_i^T x - b_i) \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \end{bmatrix} \\
&= \sum_{i=1}^{m} 2(a_i^T x - b_i) a_i \\
&= \sum_{i=1}^{m} 2(a_i a_i^T x - a_i b_i) \\
&= 2(A^T A x - A^T b)
\end{aligned}$$

Conclusão,

$$\begin{aligned}
\nabla f(x) = 0 &\Leftrightarrow AA^T x = A^T b \\
&\Leftrightarrow 2(A^T A x - A^T b) = 0 \\
&\Leftrightarrow A^T A x - A^T b = 0 \\
&\Leftrightarrow x = x_{LS}
\end{aligned}$$

Resolução Numérica do Problema de Mínimos Quadrados

Desejamos encontrar x_{LS} que é a solução do problema $A^T A x = A^T b$. Note que:

- $A^T A$ é simétrica;
- $A^T A$ é positiva semi-definida, ou seja, $x A^T A x \geqslant 0$ para todo $x \in \mathbb{R}^n$; e
- $A^T A$ é positiva definida se A tem posto completo, ou seja, $x^T A^T A x = ||Ax||^2 > 0$ para todo $x \neq 0$.

Portanto, se A tem posto completo podemos fazer uso da fatoração Cholesky para encontrar uma matriz triangular inferior L tal que $LL^T = A^T A$. Uma vez calculada a matriz Cholesky L, podemos facilmente encontrar a solução x_{LS}.

5.7 Sistemas de Equações Lineares Sub-Dimensionados

Considere o problema de encontrar uma solução para o sistema de equações lineares $Ax = b$, $A \in \mathbb{R}^{m \times n}$ e $m \leqslant n$, para o qual existem infinitas soluções. Em tal situação, pode ser útil encontrar uma solução de menor norma, matematicamente:

$$\begin{array}{ll} \text{Minimize} & \|x\| \\ \text{Sujeito a :} & \\ & Ax = b \end{array}$$

Um vetor x que resolve este problema é dito solução por norma mínima.

Teorema 5.3 *Se $rank(A) = m$, então AA^T é não singular e $x_{LN} = A^T(AA^T)^{-1}b$ é a solução única que minimiza $\|x\|$ (em outras palavras, se $Ax = b$ para $x \neq x_{LN}$, então $\|x\| > \|x_{LN}\|$).*

Prova: Primeiro, vamos mostrar que AA^T é não singular. Seja $y \in \mathbb{R}^m$ no espaço nulo de AA^T. Então:

$$\begin{array}{rcll} AA^Ty = 0 & \Rightarrow & y^TAA^Ty = 0 & \\ & \Rightarrow & \|A^Ty\|^2 = 0 & \\ & \Rightarrow & A^Ty = 0 \quad \Rightarrow \; y = 0 & [\text{pois } rank(A) = m] \end{array}$$

Segundo, vamos mostrarq que x_{LN} é uma solução:

$$Ax_{LN} = AA^T(AA^T)^{-1}b = b,$$

logo, x_{LN} é uma solução para $Ax = b$.

Terceiro, vamos mostrar que x_{LN} é uma solução ótima e única. Considere qualquer outra solução x para $Ax = b$ tal que $x \neq x_{LN}$. Então:

$$\begin{array}{rcl} (x - x_{LN})^Tx_{LN} & = & (x - x_{LN})^TA^T(AA^T)^{-1}b \\ & = & (Ax - Ax_{LN})^T(AA^T)^{-1}b \\ & = & (b - b)^T(AA^T)^{-1}b \\ & = & 0 \end{array}$$

Portanto, $(x - x_{LN})$ é ortogonal a x_{LN}. Logo,

$$\begin{array}{rcl} \|x\|^2 & = & \|x - x_{LN} + x_{LN}\|^2 \\ & = & \|x - x_{LN}\|^2 + \|x_{LN}\|^2 \\ & > & \|x_{LN}\|^2 \end{array}$$

Concluímos que x_{LN} é solução única. ■

5.7.1 Interpretação Geométrica

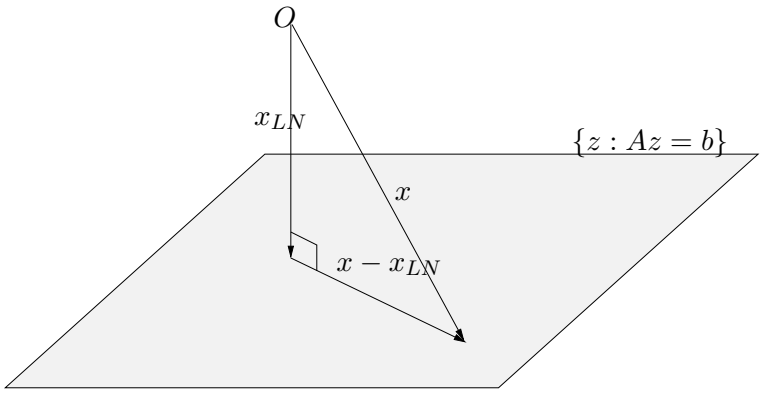


Figura 5.24: Interpretação geométrica do problema de minimização de norma

- x_{LN} é a solução de $Ax = b$ mais próxima da origem, em outras palavras, x_{LN} é a projeção da origem no espaço de soluções de $Ax = b$.
- Se $Ax = b$, então $x - x_{LN}$ é ortogonal a x_{LN}.

5.7.2 Calculando a Solução de Menor Norma

O problema então é calcular x_{LN} conforme a expressão $x_{LN} = A^T(AA^T)^{-1}b$.

Solução por meio da fatoração Cholesky

1) Obtenha $C = AA^T$
2) Obtenha uma fatoração Cholesky $C = LL^T$
3) Execute as mudanças de variáveis dadas por:

$$\begin{cases} x_{LN} &= A^T z \\ z &= (AA^T)^{-1}b \\ AA^T z &= b \end{cases} \quad \text{e} \quad \begin{cases} (LL^T)z &= b \\ Lw &= b \\ L^T z &= w \end{cases}$$

Assim o problema consiste em resolver as equações:

$$\begin{cases} x_{LN} &= A^T z \\ L^T z &= w \\ Lw &= b \end{cases}$$

4) Calcule w resolvendo por substituição a equação $Lw = b$

5) Calcule z resolvendo por substituição a equação $L^T z = w$

6) Calcule $x_{LN} = A^T z$

Exemplo

Seja o problema de minimização de norma dado por

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

Primeiro, obtemos $C = AA^T$:

$$\begin{matrix} 4 & 2 \\ 2 & \frac{3}{2} \end{matrix}$$

e calculamos a fatoração Cholesky:

$$L = \begin{bmatrix} 2 & 0 \\ 1 & \frac{1}{\sqrt{2}} \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} 4 & 2 \\ 2 & \frac{3}{2} \end{bmatrix} = AA^T = LL^T = \begin{bmatrix} 2 & 0 \\ 1 & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Resolvendo $Lw = b$, obtemos

$$\begin{bmatrix} 2 & 0 \\ 1 & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow w_1 = 0, w_2 = \sqrt{2}$$

Resolvendo $L^T z = w$, obtemos:

$$\begin{bmatrix} 2 & 1 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix} \Rightarrow z_1 = -1, z_2 = 2$$

Por fim, calculamos $x_{LN} = A^T z$:

$$x_{LN} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 1 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

5.7.3 Problemas de Minimização de Normas

Aplicações:

- sistemas de equações lineares com falta de equações
- solução por minimização de norma

5.7.4 Sistemas com Mais Variáveis do que Equações

Para sistemas subdimensionados, com mais variáveis do que equações:

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & \dots & + & a_{3n}x_n & = & b_3 \\ \vdots & + & \vdots & + & \dots & + & \vdots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{cases}$$

Ou seja, um sistema $Ax = b$ onde $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^{n \times 1}$ e $b \in \mathbb{R}^{m \times 1}$. A matriz A é gorda, $m \leqslant n$, tem mais variáveis do que equações—i.e., há várias escolhas de x que levam a uma mesma solução b.

Exemplo: Transferência de Massa

Vamos aqui considerar o problema de deslocar um bloco, conforme Figura 5.25, de um ponto a outro em um intervalo de tempo. Alguns dados do problema são:

- a massa é unitária, $m = 1Kg$;
- a velocidade e posição são nulas no instante $t = 0s$;
- o bloco está sujeito a uma força $F(t)$;
- são dados a posição e velocidade final da massa no instante $t = 10s$; e
- a força será discretizada no tempo, constante em subintervalos, tal que $F(t) = x_j$ para $j - 1 \leqslant t < j$, $j = 1, \dots, 10$.

O problema consiste em encontrar a sequência de forças $x = (x_1, \dots, x_n)$ que move a massa da origem $s(0) = 0$ até a posição final $s(10) = 1$ tendo velocidade nula na chegada, $s'(10) = 0$. Seja:

- $F(t) = x_j$ para $j - 1 \leqslant t < j$, $j = 1, \dots, 10$ a sequência de forças;
- $s(t)$ a posição da massa no instante $t = 10$; e
- $s'(t)$ a velocidade da massa no instante $t = 10$.

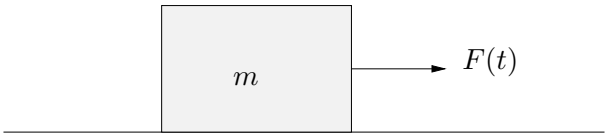


Figura 5.25: Bloco de massa

A partir da Lei de Newton, temos que no instante t a aceleração da massa é dada por $s''(t) = F(t)/m = F(t)$, pois $m = 1Kg$. Integrando a aceleração no tempo, podemos calcular a velocidade da massa no instante $t = 10$:

$$\begin{aligned}
s'(10) &= \int_0^{10} F(t)dt \\
&= \int_0^1 F(t)dt + \int_1^2 F(t)dt + \ldots + \int_9^{10} F(t)dt \\
&= \int_0^1 x_1 dt + \int_1^2 x_2 dt + \ldots + \int_9^{10} x_{10} dt \\
&= x_1 \int_0^1 dt + x_2 \int_1^2 dt + \ldots + x_{10} \int_9^{10} dt \\
&= x_1 + x_2 + \ldots + x_{10} \\
&= \sum_{i=1}^{10} x_i
\end{aligned}$$

Da mesma forma, podemos fazer uso da velocidade média em cada subintervalo para calcular a posição da massa em função do tempo:

$$\begin{aligned}
s(10) &= \int_0^{10} s'(t)dt \\
&= \int_0^1 s'(t)dt + \int_1^2 s'(t)dt + \ldots + \int_9^{10} s'(t)dt \\
&= \frac{s(1)}{2} + \left[s(1) + \frac{s(2)-s(1)}{2}\right] + \left[s(2) + \frac{s(3)-s(2)}{2}\right] + \ldots \\
&\qquad + \left[s(9) + \frac{s(10)-s(9)}{2}\right] \\
&= \frac{x_1}{2} + \left[x_1 + \frac{x_1+x_2-x_1}{2}\right] + \left[x_1+x_2 + \frac{x_1+x_2+x_3-x_1-x_2}{2}\right] + \ldots \\
&\qquad + \left[x_1+\ldots+x_9 + \frac{x_1+\ldots+x_{10}-x_1-\ldots-x_9}{2}\right] \\
&= \frac{x_1}{2} + \left(x_1 + \frac{x_2}{2}\right) + \left(x_1+x_2 + \frac{x_3}{2}\right) + \ldots + \left(x_1+\ldots+x_9 + \frac{x_{10}}{2}\right) \\
&= \left(9x_1 + \frac{x_1}{2}\right) + \left(8x_2 + \frac{x_2}{2}\right) + \left(7x_3 + \frac{x_3}{2}\right) + \ldots + \frac{x_{10}}{2} \\
&= \sum_{i=1}^{10} \left(\frac{1}{2} + 10 - i\right) x_i
\end{aligned}$$

Para o entendimento da expressão $s(10)$, o leitor pode observar a Fig. 5.26 que traz a curva da velocidade em função do tempo. No fim do primeiro intervalo, a velocidade é $s'(1) = x_1$ e, portanto, a integral da curva dada pela área do triângulo é precisamente a posição da massa, ou seja, $s(1) = x_1/2$. No fim do segundo intervalo, a velocidade é $s'(2) = x_1 + x_2$ e, portanto, a integral da curva até $t = 2$ nos dá a posição $s(2) = x_1/2 + x_1 + (x_1 + x_2 - x_1)/2 = x_1/2 + (x_1 + x_2/2)$. E assim sucessivamente.

Em notação matricial o problema pode ser colocado como o de encontrar $x \in \mathbb{R}^{10}$ tal que $Ax = b$, onde:

$$A = \begin{bmatrix} \frac{19}{2} & \frac{17}{2} & \frac{15}{2} & \cdots & \frac{1}{2} \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ e } b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Uma solução é $x_1, \ldots, x_8 = 0, x_9 = 1$ e $x_{10} = -1$. As Figuras 5.27, 5.28 e 5.29 ilustram o deslocamento $s(t)$, velocidade $s'(t)$ e força aplicada $x(t)$ ao longo do tempo. Uma segunda solução é $x_1 = 1, x_2 = -1$ e $x_3, \ldots, x_{10} = 0$. Ambas as soluções têm norma $\|x\|^2 = x_1^2 + x_2^2 + \ldots + x_{10}^2 = 2$.

A solução de menor norma tem vetor de forças ilustrado na Figura 5.30. O vetor de forças é $x_{LN} = [0.0545\ 0.0424\ 0.0303\ 0.0182\ 0.0061\ -0.0545\ -$

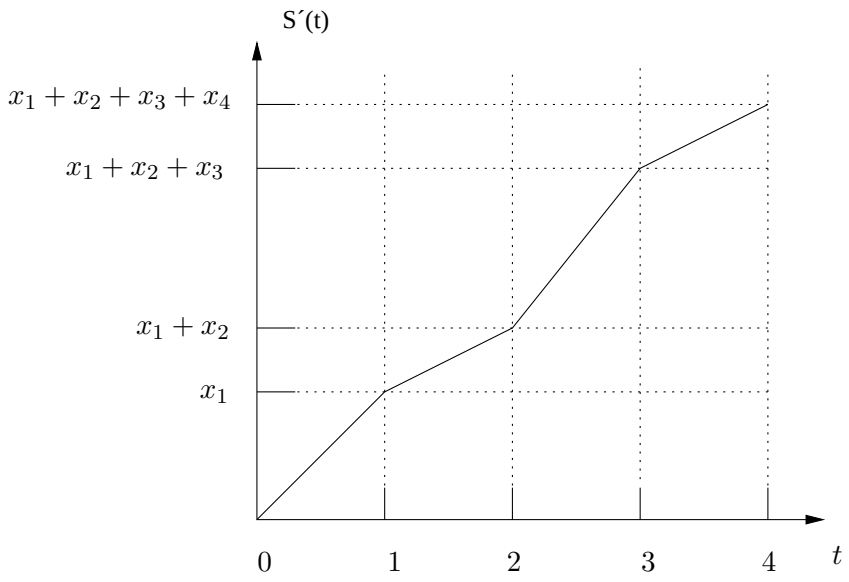


Figura 5.26: Exemplo de velocidade em função do tempo.

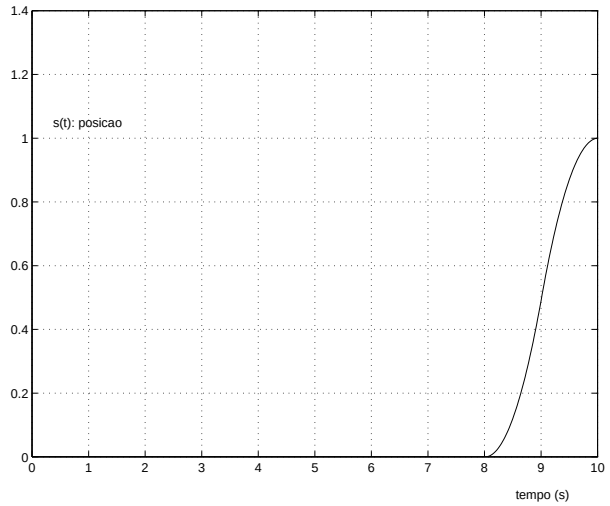


Figura 5.27: Deslocamento da massa

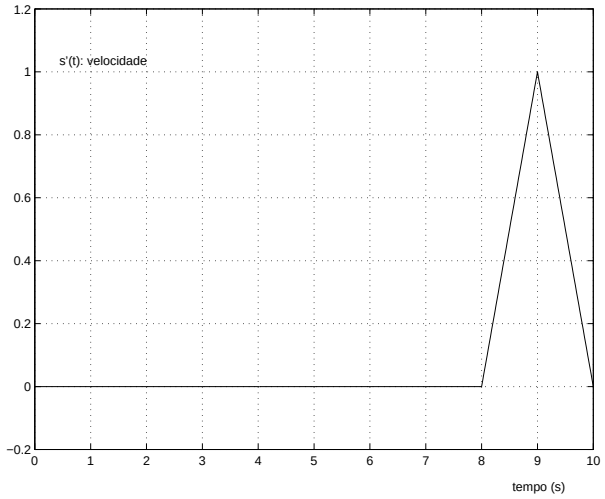


Figura 5.28: Velocidade da massa

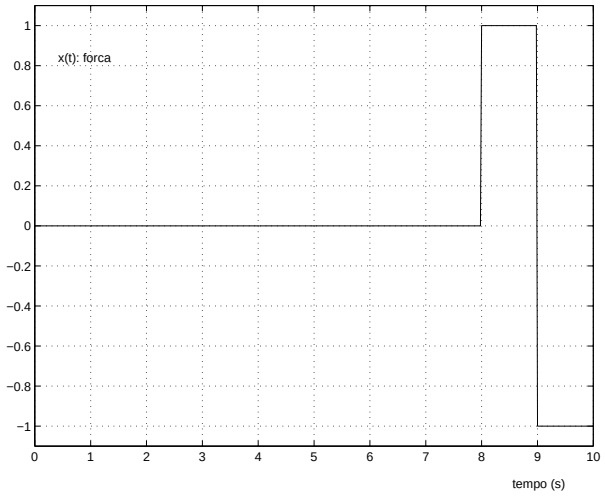


Figura 5.29: Força aplicada à massa no fim do período

$0.0424 \quad -0.0303 \quad -0.0182 \quad -0.0061]^T$, com $\|x_{LN}\| = 0.1101$ e $\|x_{LN}\|^2 = 0.0121$. O deslocamento $s(t)$ e velocidade $s'(t)$ induzidos pela força aplicada são ilustrados nas Figuras 5.31 e 5.32.

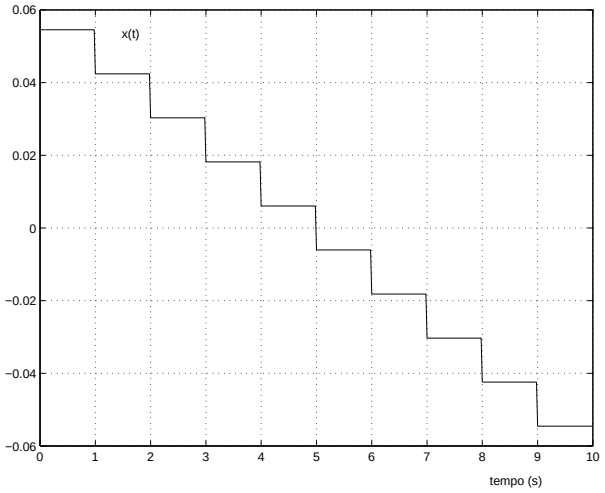


Figura 5.30: Força de menor norma aplicada à massa ao longo do intervalo

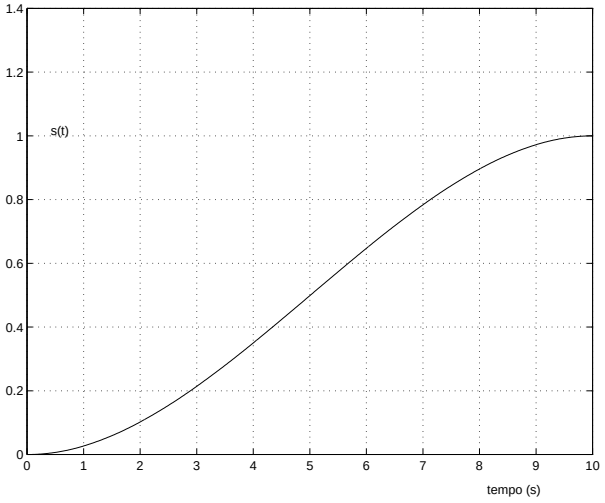


Figura 5.31: Deslocamento da massa

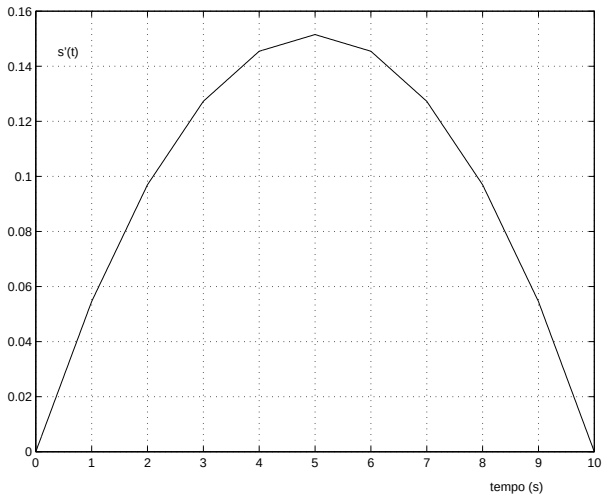


Figura 5.32: Velocidade da massa

5.8 Mínimos Quadrados Não-linear

Considere o problema de encontrar uma solução para m equações não-lineares em n variáveis:

$$\begin{cases} r_1(x_1, \ldots, x_n) = 0 \\ r_2(x_1, \ldots, x_n) = 0 \\ \qquad \vdots \\ r_m(x_1, \ldots, x_n) = 0 \end{cases}$$

onde $r_i : \mathbb{R}^n \to \mathbb{R}$ para $i = 1, \ldots, m$. O problema pode ser colocado em forma vetorial $r(x) = 0$, onde $r : \mathbb{R}^n \to \mathbb{R}^m$ é definida como:

$$r(x) = \begin{bmatrix} r_1(x_1, \ldots, x_n) \\ r_2(x_1, \ldots, x_n) \\ \vdots \\ r_m(x_1, \ldots, x_n) \end{bmatrix} \text{ e } x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Tipicamente, para $m \gg n$, não existe $x \in \mathbb{R}^n$ tal que $r(x) = 0$. Assim, nos resta buscar uma solução aproximada x tal que:

$$\begin{array}{c} \text{Minimize} \quad \|r(x)\|^2 \\ x \in \mathbb{R}^n \end{array}$$

Para o caso $r(x) = Ax - b$, o problema se reduz a um problema de mínimos quadrados linear. Note que a função $g(x) = \|r(x)\|^2$ pode e tipicamente

tem múltiplos mínimos locais. Encontrar um mínimo global é extremamente difícil, logo ficamos satisfeitos com mínimos locais de boa qualidade. Para se ter uma idéia da aplicação prática de tal problema, o problema de treinar redes Neurais pode ser visto como um caso particular do problema de mínimos quadrados não-linear.

Exemplo: Projeto de Indutor CMOS

Considere o problema de projetar simultaneamente em placa de silício um conjunto de 50 indutores. A indutância do indutor i é uma função de um vetor $x = (x_1, x_2, \ldots, x_5)$ de parâmetros de projeto, sendo dada por:

$$L_i \approx e^{x_1} n_i^{x_2} w_i^{x_3} d_i^{x_4} D_i^{x_5}$$

O problema consiste em encontrar o vetor de parâmetros $x = (x_1, x_2, \ldots, x_5)$ tal que $L_i \cong \hat{L}_i$ para $i = 1, \ldots, 50$, sendo $\hat{L}_i$ a indutância desejada para o i-ésimo indutor.

Podemos portanto colocar o problema como um problema de mínimos quadrados não-linear fazendo:

$$r_i(x) = e^{x_1} n_i^{x_2} w_i^{x_3} d_i^{x_4} D_i^{x_5} - \hat{L}_i, \ i = 1, \ldots 50$$

O problema então fica:

$$\begin{array}{ll} \text{Minimize} & \sum_{i=1}^{m} [e^{x_1} n_i^{x_2} w_i^{x_3} d_i^{x_4} D_i^{x_5} - \hat{L}_i]^2 \\ x_1, \ldots, x_5 & \end{array}$$

Será que é possível resolver o problema acima como um problema de mínimos quadrados linear?

5.9 Referências

Os tópicos sobre sistemas de equações não-lineares, mínimos quadrados, minimização de norma e introdução à otimização, apresentados neste capítulo, são sínteses das notas de aula de Vandenberghe [7].

5.10 Exercícios

Exercício 5.1 Considere os polinômios:

$$\begin{aligned} p(t) &= c_0 + c_1 t + c_2 t^2 + c_3 t^3 \\ q(t) &= d_0 + d_1 t + d_2 t^2 + d_3 t^3 \end{aligned}$$

Tal que $p(t_1) = y_1$, $p(t_2) = y_2$, $p(t_3) = y_3$, $p(t_4) = q(t_4)$, $p'(t_4) = q'(t_4)$, $q(t_5) = y_5$, $q(t_6) = y_6$ e $q(t_7) = y_7$, sendo que $t_1, \ldots, t_7$ e $y_1, \ldots, y_3, y_5, \ldots, y_7$ são dados. Formule o problema de encontrar os polinômios $p(t)$ e $q(t)$ de maneira que eles possam ser encontrados através de um método numérico. Sua formulação deve ser tão eficiente quanto possível.

Exercício 5.2 Aplique o método de Newton de forma a encontrar todas as soluções do seguinte sistema de equações não-lineares: (ilustre os passos seguidos pelo algoritmo para cada uma das raízes encontradas)

$$\begin{aligned} \log(x_1^2 + 2x_2^2 + 1) - \frac{1}{2} &= 0 \\ x_2 - x_1^2 + 0.2 &= 0 \end{aligned}$$

Exercício 5.3 Resolva o sistema de equações:

$$\begin{cases} x^2 + 2zy^3 + y^4 &= 9 \\ 3x^3 - 6x^2z - y^3 &= 4 \\ x - 2z - y &= 0 \end{cases}$$

pelo método de Newton para obter precisão até a terceira casa decimal. Use cada uma das tentativas iniciais: i) $(x_0, y_0, z_0) = (2.5, 0.5, -1)$; ii) $(x_0, y_0, z_0) = (3, 0, 2)$; iii) $(x_0, y_0, z_0) = (-3.5, -0.1, 1.2)$; e iv) $(x_0, y_0, z_0) = (3, -2, 0)$.

Exercício 5.4 Considere o problema de otimização abaixo:

$$\begin{array}{lll} P: & \textit{Minimize} & \|x\| \\ & \textit{Sujeito a}: & \\ & & Ax = b \end{array}$$

Comparando P ao problema P' de encontrarmos uma solução para $Ax = b$, sob quais condições em termos do posto da matriz A, $rank(A)$, faz sentindo resolvermos o problema P em vez de P'? Dê uma aplicação do problema P.

Exercício 5.5 Considere o problema de otimização abaixo:

$$\begin{array}{lll} \hat{P}: & \underset{x}{\textit{Minimize}} & \|Ax - b\|^2 \end{array}$$

Comparando $\hat{P}$ ao problema P' de encontrarmos uma solução para $Ax = b$, sob quais condições em termos do posto da matriz A, $rank(A)$, faz sentindo resolvermos o problema $\hat{P}$ em vez de P'? Dê uma aplicação de $\hat{P}$.

Exercício 5.6 Considere o sistema mecânico da Fig. 5.33. Há dois blocos, sendo um de massa $m_1 = 1kg$ localizado na posição $0m$ e outro de massa $m_2 = 2kg$ localizado na posição $10m$. Denota-se por x_j a força aplicada ao bloco 1 e por y_j a força aplicada ao bloco 2, ambas durante o intervalo de tempo $j-1 \leqslant t < j$ (segundos), e dadas em N. Durante o intervalo de tempo, as forças não podem ser modificadas. Seja $x = [x_1 \dots x_{10}]$ e $y = [y_1 \dots y_{10}]$ os vetores de forças para o intervalo de tempo $[0, 10s)$. Desejamos deslocar o bloco 1 até uma posição k e o bloco 2 até a posição $k + 1$, de forma que ambos fiquem adjacentes ao fim dos $10s$, sem que haja colisão.

Tarefas Específicas:

1) Encontrar os vetores de forças x e y, bem como a posição k, de forma que $f(x, y) = \|x\|^2 + \|y\|^2$ seja minimizada, tal que no instante $10s$ ambos os blocos fiquem adjacentes e com velocidade nula.

2) Simular o comportamento dos blocos para as forças aplicadas, ilustrando as forças, as velocidades e posição dos blocos em função do tempo. A simulação deverá ser feita através da integração numérica das equações diferenciais ordinárias do movimento dos objetos. Utilizar um ou mais métodos de solução numérica (e.g., Euler or Runge-Kutta). Implementar a simulação em Matlab, C, Pascal, Octave, ou outra linguagem de programação/simulação.

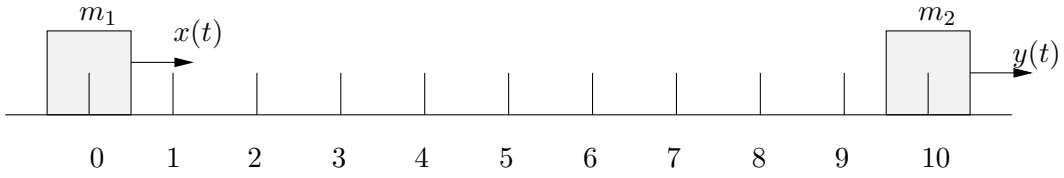


Figura 5.33: Sistema físico com dois blocos sujeitos a forças horizontais

Exercício 5.7 O problema consiste em deslocar um bloco no plano, conforme Figura 5.34 do ponto (x_0, y_0) até o ponto (x_1, y_1) em um intervalo de tempo de 20 s. Dados do problema:

- a massa é unitária, $m = 1Kg$;
- a velocidade inicial é nula em $t = 0s$;
- a posição inicial é $(x_0, y_0) = (1, -1)m$;
- o bloco está sujeito a uma força horizontal $u_x(t)$ e uma força vertical $u_y(t)$ conforme indicado na figura;

- a posição do bloco em $t = 20s$ deve ser $(x_1, y_1) = (2, 3)m$;
- a velocidade final deve ser nula; e
- a força será discretizada no tempo, constante em subintervalos de $1s$, tal que $u_x(t) = v_j^x$ e $u_y(t) = v_j^y$ para $j-1 \leqslant t < j$, $j = 1, \ldots, n$, $n = 20$.

Encontre a sequência de forças $v^x = (v_1^x, \ldots, v_n^x)$ e $v^y = (v_1^y, \ldots, v_n^y)$ que move a massa da posição $(x(0), y(0)) = (x_0, y_0)$ até o ponto $(x(20), y(20)) = (x_1, y_1)$, tendo velocidade nula na chegada, $\dot{x}(20) = 0$ e $\dot{y}(20) = 0$.

Tarefas:

1) modele o problema como um problema de minimização de norma;

2) calcule as sequências de forças que minimizam $\|(v^x, v^y)\|$ e indique o valor desta norma;

3) gere gráficos da força, posição e velocidade em função do tempo: $(v^x(t), v^y(t))$, $(x(t), y(t))$ e $(\dot{x}(t), \dot{y}(t))$.

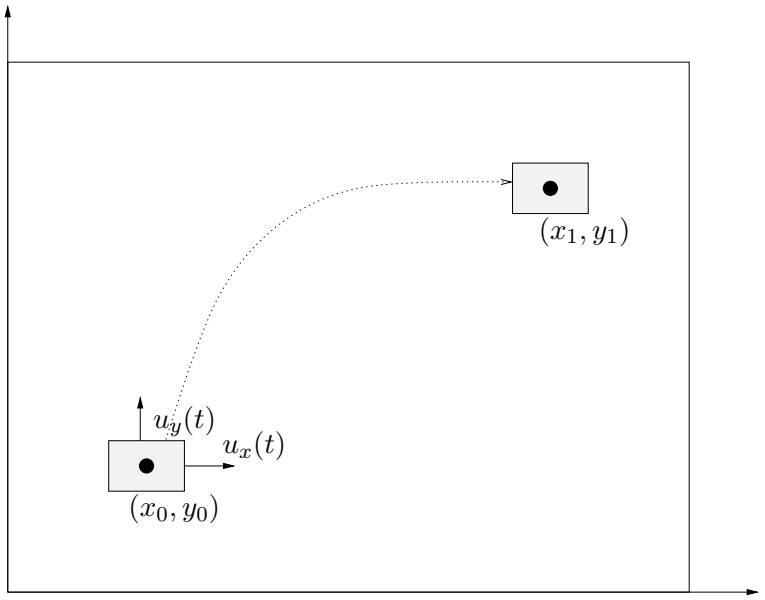


Figura 5.34: Bloco de massa

Exercício 5.8 Considere a função $f(x) = \frac{1}{2}x^3 - 10x^2 + 3x + 100$. Encontre:

a) os pontos $x \in \mathbb{R}$ que satisfazem as condições de primeira-ordem para mínimo local;

b) os pontos que satisfazem as condições necessárias de segunda-ordem para mínimo local; e

c) os pontos que satisfazem as condições suficientes de segunda-ordem para mínimo local.

Exercício 5.9 O projeto de uma família de m capacitores em circuito integrado tem como parâmetros as variáveis x, y, k, z, t. A capacitância C_j do indutor j é dada pela expressão:

$$C_j = e^x a_j^y b_j^k g_j^z d_j^t$$

onde a_j, b_j, g_j, e d_j são constantes conhecidas. Seja c_j a capacitância desejada para o j-ésimo capacitor. O problema é encontrar valores para as variáveis x, y, k, z, e t tal que $C_j = c_j$ para $j = 1, \ldots, m$. Sabemos que não há solução para o problema.

Prof. Antunes afirma que se pode resolver o problema por meio de mínimos quadrados, não sendo necessário resolver um problema de regressão não-linear. Você concorda ou discorda? Justifique.

Exercício 5.10 Seja $f : \mathbb{R}^n \to \mathbb{R}^n$ uma função não-linear contínua e diferencialmente contínua. Prof. Alberto dispõe de um pacote de otimização de funções não-lineares contínuas que resolve problemas da forma:

$$\text{Min } g(x),$$

onde $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$ e g é contínua e diferenciável. Pode-se utilizar o pacote de software para encontrar uma raiz para f, i.e., um ponto $x \in \mathbb{R}^n$ tal que $f(x) = 0$? Justifique a sua resposta.

Exercício 5.11 Questões de verdadeiro/falso.

i. Prof. Kunz afirma que não existe mínimo local para a função $f(x) = \frac{e^x}{e^x + e^{-x}}$.

ii. Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função contínua e diferenciável. Prof. Kunz afirma que se $x \in \mathbb{R}^n$ é um mínimo local, então obrigatoriamente $\nabla f(x) = 0$.

iii. Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função contínua e diferenciável. Prof. Kunz afirma que se existe um mínimo local, então deve existir um mínimo global.

iv. Seja $f : \mathbb{R} \to \mathbb{R}$ uma função contínua e duas vezes diferenciável. Para um certo $x \in \mathbb{R}$, tem-se $f'(x) = 0$ e $f''(x) \geqslant 0$. Prof. Kunz afirma que x é um mínimo local.

v. Seja $f : \mathbb{R} \to \mathbb{R}$ uma função contínua e duas vezes diferenciável tal que $f''(x) \geqslant 0$ para todo $x \in \mathbb{R}$. Prof. Kunz afirma que se x é um mínimo local, então obrigatoriamente x também é um ótimo global.

vi. Considere a função $f(x) = e^{-x}/x^2$. Dejamos encontrar o mínimo global de $f(x)$ na região $R = [1, \infty)$. Prof. Kunz diz que se $x \in R$ é um mínimo local para $f(x)$ na região R, então x é obrigatoriamente um mínimo global.

vii. Sejam $f(x, y)$ e $g(x, y)$ duas funções contínuas e diferenciáveis. Prof. Kunz afirma que se $(\hat{x}, \hat{y})$ é um mínimo local de $h(x, y) = f(x, y) + g(x, y)$ então $(\hat{x}, \hat{y})$ é também um mínimo local para f e g.

viii. Considere o problema de encontrar $x \in \mathbb{R}^n$ tal que:

$$\begin{array}{ll} \textit{Minimize} & \|x\| \\ \textit{Sujeito a}: & \\ & Ax = b \end{array}$$

onde $A \in \mathbb{R}^{m \times n}$. Prof. Kunz afirma que não faz sentido resolver este problema quando há mais equações do que variáveis, ou seja, quando $m > n$.

ix. Seja $Ax = b$ um sistema onde $A \in \mathbb{R}^{n \times n}$. Prof. Kuhn afirma que se $rank(A^T) = n$, então obrigatoriamente $b \in range(A)$ e, portanto, as soluções do problema de mínimos quadrados (x_{LS}) e do problema de minimização de norma (x_{LN}) são idênticas, ou seja, $x_{LS} = x_{LN}$.

Exercício 5.12 Seja $f : \mathbb{R} \to \mathbb{R}$ uma função não-linear contínua e diferencialmente contínua. Prof. Kunz dispõe de um pacote para solução de sistemas de equações e desigualdades não-lineares, que encontra, caso exista, uma solução x para:

$$\begin{array}{l} g(x) < 0 \\ h(x) = 0 \end{array}$$

onde $g : \mathbb{R} \to \mathbb{R}$ e $h : \mathbb{R} \to \mathbb{R}$ são funções contínuas e diferenciáveis. Prof. Kunz afirma que se pode utilizar o pacote de software para encontrar um mínimo local para $f(x)$. Você concorda ou discorda do Prof. Kunz? Se você discorda, mostre porque não podemos utilizar o pacote de software. Se você concorda, mostre como que um mínimo local pode ser encontrado.

Exercício 5.13 Considere o sistema de equações lineares dado por:

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} x = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \equiv Bx = b \tag{5.9}$$

onde $B_1 \in \mathbb{R}^{m_1 \times n}$, $B_2 \in \mathbb{R}^{m_2 \times n}$, $B \in \mathbb{R}^{(m_1+m_2)\times n}$ e os vetores x, b_1, b_2 e b têm dimensões apropriadas.

Sabe-se que o sistema (5.9) não possui solução, mas o sub-sistema $B_1 x = b_1$ possui infinitas soluções. Deseja-se resolver o seguinte problema de mínimos quadrados sob restrições:

$$\begin{aligned} &\textit{Minimize} && \|Bx - b\|^2 && \text{(5.10a)} \\ &\textit{Sujeito a:} && && \text{(5.10b)} \\ & && B_1 x = b_1 && \text{(5.10c)} \end{aligned}$$

É possível resolver esta generalização do problema de mínimos quadrados usando apenas os modelos e algoritmos desenvolvidos na disciplina? Se sim, mostre como a solução pode ser obtida. Caso contrário, argumente uma justificativa da impossibilidade.

Exercício 5.14 A aplicação do método de Newton para determinar uma raiz da função $f(x) = x^3 - 2x + 2$ pode entrar em laço infinito.

i. Obtenha uma função $F : \mathbb{R}^2 \to \mathbb{R}^2$ tal que toda a solução do sistema de equações não lineares $F(y) = 0$ define um par de pontos, $x = y_1$ e $\bar{x} = y_2$, para os quais o método de Newton entra em laço infinito quando aplicado à equação $f(x) = 0$.

ii. Aplique o método de Newton ao sistema $F(x) = 0$ e obtenha uma solução $y = (y_1, y_2)$.

Capítulo 6

Revisão de Polinômios

Definição 6.1 *Um polinômio p é uma função com domínio e imagem em um conjunto $\mathbb{C}$ ou $\mathbb{R}$ dado na forma:*

$$\begin{array}{rccl} p: & \mathbb{C} & \to & \mathbb{C} \\ & x & \mapsto & p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n \end{array}$$

O número inteiro n é dito grau do polinômio.

Teorema 6.1 *Se $p(x)$ é um polinômio de grau n, então para qualquer α existe um polinômio $q(x)$ único tal que: $p(x) = (x-\alpha)q(x) + p(\alpha)$*

O Teorema 6.1 nos diz que, se dividirmos $p(x)$ por $(x-\alpha)$ então encontramos como quociente um polinômio de grau $n-1$, se $n > 1$, e o resto é o valor do polinômio calculado em α.

Exemplo

Considere o polinômio $p(x) = 3x^5 + 4x^4 - 2x^3 - x^2 + 3x - 4$ de grau 5. Para calcularmos o valor do polinômino para $x = 2$, $p(2)$, podemos fazer as seguintes contas:

$$p(2) = 3 \times 2^5 + 4 \times 2^4 - 2 \times 2^3 - 2^2 + 3 \times 2 - 4$$

o que implica em executarmos n adições e $\sum_{j=1}^{n} j = \frac{(n+1)n}{2}$ multiplicações. Portanto, o procedimento executa $\Theta(n)$ adições e $\Theta(n^2)$ multiplicações, o que nos leva a concluir que a complexidade computacional do procedimento acima é da ordem $\Theta(n^2)$ operações computacionais elementares. Será que este é o procedimento mais eficiente?

Entretanto, observamos que:

$$\begin{aligned} p(x) &= 3x^5 + 4x^4 - 2x^3 - x^2 + 3x - 4 \\ &= (3x^4 + 4x^3 - 2x^2 - x + 3)x - 4 \\ &= ((3x^3 + 4x^2 - 2x - 1)x + 3)x - 4 \\ &= (((3x^2 + 4x - 2)x - 1)x + 3)x - 4 \\ &= ((((3x + 4)x - 2)x - 1)x + 3)x - 4 \end{aligned}$$

o que resulta no cálculo de $p(2)$ com apenas n adições e n multiplicações. Este segundo procedimento é muito mais eficiente que o anterior, tem uma complexidade computacional de $\Theta(n)$ operações.

O esquema de cálculo de $p(x)$ acima pode ser utilizado para dividirmos $p(x)$ por $(x - \alpha)$ e daí calculamos $q(x) = b_0x^4 + b_1x^3 + b_2x^2 + b_3x + b_4$ e $p(\alpha)$.

Esquema de Horner/Briot-Ruffini

O esquema para calcular o quociente é ilustrado na tabela abaixo, onde estamos buscando o quociente $q(x)$ da divisão do polinômio $p(x) = 3x^5 + 4x^4 - 2x^3 - x^2 + 3x - 4$ por $(x - \alpha)$, com $\alpha = 2$.

	3	4	-2	-1	3	-4
$\alpha = 2$		6	20	36	70	146
	3	10	18	35	73	142
	b_0	b_1	b_2	b_3	b_4	$R = p(2)$

A partir da tabela acima, obtemos o quociente através dos coeficientes da linha mais abaixo, ou seja, $q(x) = 3x^4 + 10x^3 + 18x^2 + 35x + 73$ e também o resto $p(2) = 142$. Portanto, $p(x) = (x - 2)q(x) + 142$.

Na forma mais geral, o método Briot-Ruffini pode ser expresso através das operações indicadas na tabela a seguir:

	a_0	a_1	a_2	a_3	$\ldots$	a_{n-1}	a_n
α		$b_0\alpha$	$b_1\alpha$	$b_2\alpha$	$\ldots$	$b_{n-2}\alpha$	$b_{n-1}\alpha$
	a_0	$a_1 + b_0\alpha$	$a_2 + b_1\alpha$	$a_3 + b_2\alpha$	$\vdots$	$a_{n-1} + b_{n-2}\alpha$	$a_n + b_{n-1}\alpha$
	b_0	b_1	b_2	b_3		b_{n-1}	$R = p(\alpha)$

Corolário 6.1 *Se $p(x)$ é um polinômio de grau $n > 1$ e $p(\alpha) = 0$ então existe um polinômio único de grau $n - 1$, tal que $p(x) = (x - \alpha)q(x)$. Neste caso, $q(x)$ é chamado de polinômio reduzido.*

6.1 Enumeração de Raízes

6.1.1 Enumeração das Raízes de Uma Equação Polinomial

Enumerar as raízes de um polinômio $p(x)$ consiste em dizermos quantas raízes o polinômio possui e de que tipo elas são. No que segue são apresentados alguns teoremas e outros resultados teóricos que podem nos auxiliar na tarefa de enumeração.

Teorema 6.2 *O número de raízes positivas de uma equação polinomial $p(x)$ com coeficientes reais, nunca é maior que o número de trocas de sinal T na sequência de seus coeficientes não nulos, e se é menor, então é sempre por um número par.*

Exemplo

Como exemplo, tome o polinômio $p(x) = x^3+2x^2-3x-5$, o qual apresenta a sequência de sinais $(+,+,-,-)$. Logo, segundo o Teorema 6.2, $T' = 1$ e pode-se afirmar com exatidão que $p(x)$ tem uma raiz positiva já que ele não pode ter um número negativo de raízes.

Observação: A mesma regra acima, dada pelo Teorema 6.2, pode ser aplicada para a enumeração das raízes reais e negativas de $p(x)$, calculando-se $p(-x)$, pois as raízes positivas de:

$$p(-x) = -x^3 + 2x^2 + 3x - 5$$

se referem às raízes negativas de $p(-x)$. Notando que a sequência de sinais de $p(-x)$ é $(-,+,+,-)$, concluímos que $T' = 2$ e daí deduzimos que $p(x)$ pode ter duas ou zero raízes negativas. Tomando como base as deduções de que $p(x)$ tem uma raiz positiva e duas ou nenhuma raiz negativa, podemos deduzir que:

- Se $p(x)$ tiver duas raízes negativas, então não terá nenhuma raiz complexa. Se, contudo, não tiver raízes negativas, então terá duas complexas.

- É bom lembrar que, se um polinômio tem todos os coeficientes reais e se houver uma raíz complexa, então sua conjugada, também será raiz do polinômio.

Exemplo

Seja $p(x) = x^4 - x^3 + x^2 - x + 1$ um polinômio de quarto grau. Temos que $T = 4$ e, portanto, $p(x)$ tem quatro, duas ou não tem raízes positivas. Procedendo à análise de $p(-x) = x^4 + x^3 + x^2 + x + 1$, observamos que $T' = 0$ e daí verificamos que $p(x)$ não tem raízes negativas. Logo $p(x)$ pode ter quatro raízes positivas, ou duas raízes positivas e duas complexas, ou nenhuma positiva e quatro complexas. Há apenas três possibilidades quanto aos tipos das raízes.

6.1.2 Enumeração das Raízes Complexas

Nesta seção damos continuidade a formas e métodos de se enumerar raízes, onde serão enunciados resultados teóricos e procedimentos de enumeração.

Teorema 6.3 *(Regra de du Gua) Dada a equação polinomial $p(x) = 0$ de grau n sem raízes nulas e se para algum k, $1 \leqslant k < n$ tivermos $a_k^2 \leqslant a_{k+1}a_{k-1}$ então $p(x)$ terá raízes complexas.*

O Teorema 6.3 nos dá condições suficientes para existência de raízes complexas. Note que se as condições do teorema não puderem ser aplicadas, o polinômio pode ter raízes complexas. A regra da Lacuna abaixo enunciada permite a conclusão sobre a existência de raízes complexas

Teorema 6.4 *(Regra da Lacuna)*

- *Se os coeficientes de $p(x)$ forem todos reais e para algum k, $1 \leqslant k < n$ tivermos $a_k = 0$ e $a_{k+1}a_{k-1} > 0$, então $p(x) = 0$ terá raízes complexas.*
- *Se os coeficientes forem todos reais e existirem dois ou mais coeficientes nulos sucessivos, então $p(x) = 0$ terá raízes complexas.*

Exemplo

Vamos agora exemplificar a aplicação dos teoremas enunciados. Inicialmente, tomemos o polinômio $p(x) = 2x^5 + 3x^4 + x^3 + 2x^2 - 5x + 3$, para o qual verificamos que $T = 2$, e daí descobrimos que $p(x)$ tem duas raízes ou zero raízes positivas. A partir de $p(-x) = -2x^5 + 3x^4 - x^3 + 2x^2 + 5x + 3$, calculamos que $T' = 3$, e daí deduzimos que $p(x)$ tem três raízes ou uma raíz real negativa. Na tabela abaixo listamos todas as possíveis combinações de tipos de raízes.

Reais Positivas	Reais Negativas	Complexas	Total
2	3	0	5
2	1	2	5
0	3	2	5
0	1	4	5

Pela regra de "du Gua" (ver Teorema 6.3), temos que ${a_2}^2 \leqslant a_3 a_1 \Rightarrow 1 \leqslant 3 \times 2 = 6$. Daí chegamos à conclusão que $p(x)$ tem raízes complexas e, por conseguinte, podemos eliminar a primeira alternativa do quadro anterior, restando apenas três possibilidades para as raízes.

Exemplo

Repetindo os passos anteriores, tomemos agora o polinômio de sexto grau $p(x) = 2x^6 - 3x^5 - 2x^3 + x^2 - x + 1$. A partir do fato que $T = 4$, concluímos que $p(x)$ tem quatro, ou duas raízes, ou zero raízes positivas. Através do polinômio $p(-x) = 2x^6 + 3x^5 + 2x^3 + x^2 + x + 1$, temos que $T' = 0$, portanto, $p(x)$ não tem raízes reais negativas. Pela Regra da Lacuna temos que $p(x) = 0$ tem raízes complexas pois: $a_2 = 0$ e $a_1 a_3 > 0$. Os possíveis arranjos de tipos e números de raízes é dado no quadro abaixo:

Reais Positivas	Reais Negativas	Complexas	Total
4	0	2	6
2	0	4	6
0	0	6	6

Definição 6.2 *Seja $f(x) = 0$ uma equação onde $f : \mathbb{R} \mapsto \mathbb{R}$ é uma função qualquer. Se $f(\bar{x}) = 0$, então dizemos que $\bar{x}$ é uma raiz de f.*

Definição 6.3 *Se $\bar{x}$ é um zero de $f(x)$ então a multiplicidade m de $\bar{x}$ é o ínfimo de todos os números k, tais que:*

$$\lim_{x \to \bar{x}} \frac{|f(x)|}{|x - \bar{x}|^k} < \infty$$

Exemplo

Consideremos a função $f(x) = x^{\frac{1}{2}}$, uma raiz de $x^{\frac{1}{2}} = 0$ é $\bar{x} = 0$. Esta raiz tem multiplicidade $\frac{1}{2}$, pois

$$\lim_{x \to 0} \frac{|x^{\frac{1}{2}}|}{|x|^{\frac{1}{2}}} < \infty \text{ mas } \lim_{x \to 0} \frac{|x^{\frac{1}{2}}|}{|x|^a} = \infty \text{ para } a < \frac{1}{2}.$$

Teorema 6.5 *Se $\bar{x}$ é um zero de f e se para algum inteiro m, $f(x)$ é m vezes continuamente diferenciável, então a multiplicidade de $\bar{x}$ é pelo menos m vezes se, e somente se,*

$$f(\bar{x}) = f'(\bar{x}) = f''(\bar{x}) = \ldots = f^{m-1}(\bar{x}) = 0$$

A multiplicidade é exatamente m se $f^m(\bar{x}) \neq 0$

Teorema 6.6 *Seja $p(x)$ um polinômio de grau $n > 1$. A multiplicidade de um zero α de $p(x)$ é m se, e somente se,*

$$\begin{cases} p(\alpha) = p'(\alpha) = p''(\alpha) = \ldots = p^{m-1}(\alpha) = 0 \\ p^m(\alpha) \neq 0. \end{cases}$$

Teorema 6.7 *Seja $p(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$, um polinômio de grau n. Então existem números distintos $\alpha_1, \alpha_2, \ldots, \alpha_s$ (que podem ser complexos) e inteiros $m_1, m_2, \ldots, m_s$ tal que para uma constante c única temos:*

$$p(x) = c(x - \alpha_1)^{m_1}.(x - \alpha_2)^{m_2} \ldots \ (x - \alpha_s)^{m_s}$$

$$\sum_{j=1}^{s} m_j = n.$$

O teorema acima é decorrência do teorema fundamental da Álgebra, que diz que todo polinômio com coeficientes complexos admite pelo menos uma raiz complexa. Nem todo o polinômio real admite uma raíz real, por exemplo $x^2 + 1$ só admite raízes complexas.

Teorema 6.8 *Se os coeficientes de $p(x)$ são reais e μ é a multiplicidade de uma raiz α então perto de α o polinômio $p(x)$ deve ter uma das formas da Figura 6.1.*

Observações:

- A enumeração de raízes reais ou complexas pode ser feita aproximadamente pelo método gráfico a ser visto posteriormente.
- A existência de um máximo local negativo, ou mínimo local positivo indica a existência nas proximidades de raízes complexas.

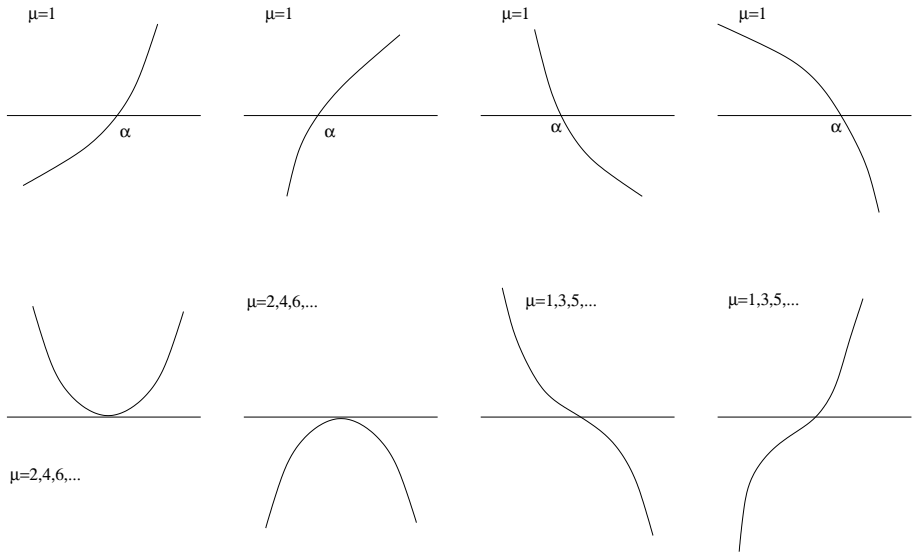


Figura 6.1: Multiplicidade de raízes

6.2 Localização das Raízes

6.2.1 Localização das Raízes Reais de Uma Equação Polinomial

Dada a equação polinomial $p(x) = 0$ podemos ter uma idéia mais ou menos precisa sobre quantos e de que tipo são as raízes de equação polinomial. Este tópico foi objeto de estudo na seção anterior. É preciso também saber onde elas estão localizadas, o que será o foco na presente seção. Serão apresentadas definições e teoremas que permitem realizar a localização das raízes.

Definição 6.4 *Localizar as raízes de $p(x) = 0$ é determinar um intervalo que contenha todas as raízes reais de $p(x)$. Localizar as raízes complexas é determinar os raios interno e externo que contenham as raízes complexas de $p(x) = 0$.*

A Figura 6.2 ilustra o conceito de localização de raízes reais e complexas.

Teorema 6.9 *(Laguerre) Dado o polinômio $p(x)$ de coeficientes reais e dado um número α, obtemos $p(x) = (x - \alpha)q(x) + R$. Se os coeficientes de $q(x)$ e R forem todos positivos ou nulos, então teremos que todas as raízes reais positivas x_j verificam $x_j < \alpha$.*

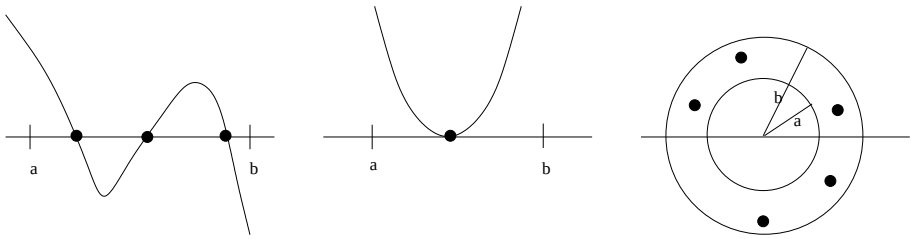


Figura 6.2: Localização de raízes

Cota de Laguerre-Thibault

Dado $p(x) = 0$ de coeficientes reais, faça a deflação de $p(x)$ por $x-1$, $x-2$, $x-3\ldots$, até $x-m$, onde $q(x)$ tenha todos os coeficientes positivos ou nulos, assim como $R(x) > 0$ tal m é chamado de cota superior das raízes reais de $p(x) = 0$. Para determinar a cota inferior basta fazer o mesmo procedimento para $p(-x)$ e assim determina-se a cota inferior.

Exemplo

Tomemos como exemplo o polinômio $p(x) = x^5 + x^4 - 9x^3 - x^2 + 20x - 12$ e considere a tarefa de localizarmos as raízes de $p(x) = 0$.

	1	1	-9	-1	20	-12
1		1	2	-7	-8	12
	1	2	-7	-8	-12	0

	1	1	-9	-1	20	-12
2		2	6	-6	-14	12
	1	3	-3	-7	6	0

	1	1	-9	-1	20	-12
3		3	12	9	24	132
	1	4	3	8	44	120

Logo temos que todas as raízes positivas de $p(x) = 0$ são menores que 3. Para pesquisar a localização das raízes negativas utiliza-se o mesmo procedimento, mas desta vez este é aplicado ao polinômio obtido ao multiplicar-se $p(-x) = -x^5 + x^4 + 9x^3 - x^2 - 20x - 12$ por -1.

	1	-1	-9	1	20	12
1		1	0	-9	-8	12
	1	0	-9	-8	12	24

	1	-1	-9	1	20	12
2		2	2	-14	-26	-12
	1	1	-7	-13	-6	0

	1	-1	-9	1	20	12
3		3	6	-9	-24	-12
	1	2	-3	-8	-4	0

	1	-1	-9	1	20	12
4		4	12	12	52	288
	1	3	3	13	72	300

Portanto, as raízes pertencem ao intervalo [-4,3].

Teorema 6.10 *(Cota de Vene) Para toda a raiz positiva α de $p(x) = 0$ onde $a_0 \neq 0$, verifica-se que:*

$$0 \leqslant \alpha \leqslant 1 + \frac{M}{a_0 + a_1 + \ldots + a_p}$$

onde M é o valor absoluto do menor dos coeficientes negativos e a_p é o ultimo coeficiente positivo antes do primeiro coeficiente negativo.

Exemplo

Vamos ilustrar a Cota de Vene com o polinômio $p(x) = x^5 + x^4 - 9x^3 - x^2 + 20x - 12$. Note que $M = |-12| = 12$ e $a_p = 1$. Logo, $0 \leqslant \alpha \leqslant 1 + \frac{|-12|}{1+1} = 1 + \frac{12}{2} = 7$.

6.3 Localização das Raízes Complexas de Uma Equação Polinomial

Teorema 6.11 *(Cota de Kojima) Dado o polinômio*

$$p(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

toda a raiz α, real ou complexa, verifica que

$$|\alpha| \leqslant q_1 + q_2$$

onde q_1 e q_2 são os valores maiores de:

$$\left\{ \left| \frac{a_i}{a_0} \right|^{\frac{1}{i}} \right\}, \ i = 1, 2, \ldots, n.$$

Exemplo

Seja o polinômio $p(x) = x^5 + x^4 - 9x^3 - x^2 + 20x - 12$, onde $a_0 = 1$, $a_1 = 1$, $a_2 = -9$, $a_3 = -1$, $a_4 = 20$, e $a_5 = -12$. Podemos então verificar que a série de fatores é:

$$\left\{ 1^{\frac{1}{1}}, 9^{\frac{1}{2}}, 1^{\frac{1}{3}}, 20^{\frac{1}{4}}, 12^{\frac{1}{5}} \right\} = \{1, 3, 1, 2.114743537, 1.643751829\}.$$

Daí verificamos que $q_1 = 3$ e $q_2 = 2.114742527$. Logo, temos que toda raiz satisfaz $|\alpha| \leqslant 5.114742527$, o que nos dá o raio externo do anel que contém as raízes de $p(x)$.

Para determinar o raio interno do anel, devemos calcular $p(\frac{1}{x}) = -12x^5 + 20x^4 - x^3 - 9x^2 + x + 1$ e aplicar o mesmo procedimento, pois as raízes de $p(\frac{1}{x})$ são os inversos das de $p(x)$. Temos então que:

$$\left\{ \left(\frac{20}{12}\right)^{\frac{1}{1}}, \left(\frac{1}{12}\right)^{\frac{1}{2}}, \left(\frac{9}{12}\right)^{\frac{1}{3}}, \left(\frac{1}{12}\right)^{\frac{1}{4}}, \left(\frac{1}{12}\right)^{\frac{1}{5}} \right\} =$$

$$\{1.666, 0.288675239, 0.908560296, 0.537284, 0.608364342\}.$$

Daí verificamos que $q_1 = 1.666$ e $q_2 = 0.908560296$. $c = 2.575226902$ e daí a cota inteira é $\frac{1}{c} = 0.388315288 \Rightarrow 0.388315288 < |\alpha| < 5.114742527$.

Teorema 6.12 *(Cota de Cauchy) Dado um polinômio real $p(x)$, então toda raiz α real ou complexa, de $p(x) = 0$ satisfaz:*

$$|\alpha| < |\beta|$$

sendo $\beta = \lim_{i \to \infty} x_i$ com $x_0 = 0$ e

$$x_i = \left(|\frac{a_1}{a_0}| x_{i-1}^{n-1} + |\frac{a_2}{a_0}| x_{i-1}^{n-2} + \ldots + |\frac{a_{n-1}}{a_0}| x_{i-1} + |\frac{a_n}{a_0}| \right)^{\frac{1}{n}}.$$

Exemplo

Tomemos o polinômio $p(x) = x^5 + x^4 - 9x^3 - x^2 + 20x - 12$. Então $x_0 = 0$ e $a_0 = 1$, o que nos leva a produzir a série:

$$x_{k+1} = [x_k^4 + 9x_k^3 + x_k^2 + 20x_k + 12]^{\frac{1}{5}}$$

Podemos então verificar que

$$\begin{aligned} x_1 &= \sqrt[5]{12} = 1.643751829 \\ x_2 &= 2.485458195 \\ &\vdots \\ x_{10} &= 3.805140857 \end{aligned}$$

Como resultado, podemos afirmar que não há nenhuma raiz fora do círculo centrado em zero e de raio 4.0.

6.4 Separação das Raízes de Uma Equação Polinomial

Definição 6.5 *Separar as raízes de uma equação polinomial é o processo de encontrar uma sequência de subintervalos distintos, tais que cada subintervalo contenha exatamente uma raiz real e cada raiz real esteja num subintervalo.*

Teorema 6.13 *(Bolzano) Se f for uma função continua em $[a, b]$ e trocar de sinal nos extremos desse intervalo, então existe pelo menos uma raiz real de f em $[a, b]$.*

Teorema 6.14 *(Budan) Seja $p^k(a)$ o valor da derivada de ordem k de $p(x)$ calculada para $x = a$. Seja v_a o número de variações de sinal da sequência:*

$$p(a), p'(a), p''(a), \ldots, p^{(n+1)}(a)$$

tomadas nesta ordem. Então o número de raízes de $p(x) = 0$ no intervalo (a, b) é igual ou menor que $|v_a - v_b|$ por um múltiplo de 2.

Exemplo

Seja $p(x) = x^3 - 2x^2 - x + 2$ um polinômio de grau 3. Pela regra de Descartes temos duas variações de sinal e daí segue que existem duas raízes,

ou não temos raízes positivas. Vamos agora calcular a forma analítica das derivadas de $p(x)$:

$$\begin{aligned} p'(x) &= 3x^2 - 4x - 1 \\ p''(x) &= 6x - 4 \\ p'''(x) &= 6 \end{aligned}$$

Calculando a cota de Laguerre-Thibault, conforme tabelas que seguem abaixo, podemos deduzir que as raízes positivas são menores que 3 (cota superior igual a 3).

	1	-2	-1	2
1		1	-1	-2
	1	-1	-2	0

	1	-2	-1	2
2		2	0	-2
	1	0	-1	0

	1	-2	-1	2
3		3	3	6
	1	1	12	8

Aplicando o Teorema de Budan, temos que $v_0 = 2$ e $v_3 = 0$, conforme tabelas abaixo, logo há duas ou nenhuma raiz real em $[0, 3]$.

$$\begin{array}{rcr|rcr} p(0) &=& 2 & p(3) &=& 8 \\ p'(0) &=& -1 & p'(3) &=& 10 \\ p''(0) &=& -4 & p''(3) &=& 14 \\ p'''(0) &=& 6 & p'''(3) &=& 6 \end{array}$$

Exemplo

Considere agora o polinômio:

$$p(x) = x^4 - 4x^3 + 6x^2 + 4x + 1$$

de grau 4. O número de variações de sinal de $p(x)$ é 4, donde podemos ter quatro, duas ou nenhuma raiz positiva. Vamos então calcular a Cota de Laguerre-Thibault, conforme desenvolvimento abaixo.

	1	-4	6	4	1
1		1	-3	3	7
	1	-3	3	7	8

	1	-4	6	4	1
2		2	-2	8	24
	1	-2	4	12	25

	1	-4	6	4	1
3		3	-3	3	21
	1	-1	3	7	22

	1	-4	6	4	1
4		4	0	24	56
	1	0	6	28	57

Uma vez que a Cota de Laguerre-Thibault é 4, podemos aplicar o Teorema de Budan calculando v_0 e v_4, mas antes disso temos que obter as derivadas de $p(x)$:

$$\begin{aligned} p'(x) &= 4x^3 - 12x^2 + 12x + 4 \\ p''(x) &= 12x^2 - 24x + 12 \\ p'''(x) &= 24x - 24 \\ p''''(x) &= 24 \end{aligned}$$

Daí calculamos os valores das derivadas nos pontos $x = 0$ e $x = 4$, obtendo os resultados do quadro abaixo.

$$\begin{array}{rcr|rcr} p(0) &=& 1 & p(4) &=& 81 \\ p'(0) &=& 4 & p'(4) &=& 108 \\ p''(0) &=& 12 & p''(4) &=& 108 \\ p'''(0) &=& -24 & p'''(4) &=& 72 \\ p''''(0) &=& 24 & p''''(4) &=& 24 \end{array}$$

Portanto, deduzimos que $v_0 = 2$ e $v_4 = 0$. O teorema então nos diz que o número de raízes em $(0, 4)$ é menor ou igual a $|v_0 - v_4| = 2$ por um múltiplo de 2. Portanto, deve haver duas ou nenhuma raiz no intervalo $(0, 4)$.

6.5 Exercícios

Exercício 6.1 Considere o polinômio:

$$p(x) = x^7 + 4x^6 - 7x^5 - 34x^4 - 24x^3 - x + 1$$

a) Prof. Miguel afirma que $p(x)$ deve obrigatoriamente possuir pelo menos uma raiz real negativa e pelo menos duas raízes complexas. Você concorda ou discorda da afirmação? Justifique a resposta.

b) Prof. Clóvis afirma que toda raiz α de $p(x)$ verifica $|\alpha| \leqslant 7.24$. Você concorda ou discorda do Prof. Clóvis? Justifique a resposta.

Exercício 6.2 Considere o polinômio:

$$p(x) = 2x^7 + 4x^6 - 7x^5 + 12x^4 - x - 1$$

Para cada afirmação, indique se você concorda ou discorda e justifique a sua resposta.

a) Prof. Miguel afirma que $p(x)$ deve obrigatoriamente possuir pelo menos uma raiz real positiva.

b) Prof. Julius afirma que $p(x)$ possui pelo menos duas raízes reais negativas.

c) Prof. Antunes afirma que $p(x)$ possui pelo menos duas raízes complexas.

d) Prof. Albus afirma que, se $p(x)$ possuir uma raiz complexa, então se α é uma raiz complexa, ela satisfaz a relação: $|\alpha| \leqslant 4$.

Exercício 6.3 Dado o polinômio:

$$q(x) = 2x^7 + 6x^6 - 10x^5 - 30x^4 + 8x^3 + 24x^2,$$

execute as tarefas abaixo dando justificativas.

a) Encontre os números possíveis de raízes reais positivas.

b) Encontre os números possíveis de raízes reais negativas.

c) Podemos dizer que $q(x)$ possui raízes complexas?

d) Se $q(x)$ possui raízes reais, então localize estas raízes.

e) Se $q(x)$ possui raízes complexas, então localize estas raízes.

Capítulo 7

Integração Numérica

Ao contrário da diferenciação, a integral de uma função $f(x)$ não necessariamente possui uma solução analítica. Por exemplo, a integral limitada $F(x) = \int_a^b e^{-x^2} dx$ da função $f(x) = e^{-x^2}$ não possui solução analítica. Então, como podemos encontrar $F(x)$? Uma solução aproximada (arbitrariamente aproximada para o caso de uma máquina de precisão ilimitada) pode ser obtida por meio de métodos numéricos. Este métodos serão objeto de estudo no presente capítulo.

7.1 O Problema da Integração Numérica

Os métodos para o cálculo de integrais definidas $F(x) = \int_a^b f(x)dx$ são agrupadas em quatro tipos:

Método Analítico: Este método consiste em se encontrar a solução analítica de $F(x)$, por exemplo $F(x) = \int \frac{1}{x} dx = \ln x + c$. Por outro lado, $F(x) = \int e^{-x^2} dx$ não pode ser escrita como uma combinação finita de outras funções algébricas, logarítmicas ou exponenciais. No caso de $F(x) = \int_0^x \frac{1}{1+x^8} dx$, podemos obter $F(x)$ através de várias etapas mas estas podem levar a erros e, além disso, o resultado pode envolver algumas funções que serão avaliadas numericamente que, por sua vez, poderiam acarretar erros numéricos. Há também ferramentas computacionais inspiradas em algoritmos de Inteligência Artificial (IA) que encontram as primitivas de várias funções, tais como as ferramentas encontradas em pacotes de software como Mathematica, Maple e Matlab.

Método Mecânico: Tais métodos fazem uso de instrumentos que calculam a área delimitada por uma curva qualquer, todavia são limitados quanto

ao número de dimensões e têm aplicações restritas.

Método Gráfico: Toma como base o desenho de $y = f(x)$ no intervalo $[a, b]$ que gera uma seqüência de iterandos no gráfico até que se obtenha o resultado. Estes métodos são pouco empregados uma vez que não são automáticos e portanto não podem ser aplicados em sistemas computacionais.

Método Numérico ou Algorítmico: Os métodos numéricos podem ser empregados em geral e tem grande apelo prático uma vez que podem ser embutidos em ambientes computacionais.

7.2 Objetivo da Integração Numérica

O método numérico para calcular a integral de $f(x)$ utiliza exclusivamente as operações aritméticas necessárias ao cálculo de $f(x)$, o que pode ser conveniente, assim dispensando o cômputo das derivadas de f. Usualmente vamos calcular a integral de $f(x)$ de a até b, ou seja, $F(x) = \int_a^b f(x)dx$, onde $-\infty < a < b < +\infty$.

7.2.1 Filosofias Básicas

Para calcular o valor aproximado da integral definida vamos utilizar uma combinação linear de valores da função $f(x)$ em certos pontos x_j, $a \leqslant x_j \leqslant b$, chamados de nós e certos valores w_j que constituem os pesos. Mais formalmente, vamos aproximar $F(x)$ com a expressão:

$$\int_a^b f(x)dx \cong w_1 f(x_1) + w_2 f(x_2) + \ldots + w_{n+1} f(x_{n+1}) = \sum_{j=1}^{n+1} w_j f(x_j) \quad (7.1)$$

De acordo com os valores dos pesos e com a escolha de nós, temos no lado direito de (7.1) o que chamamos de *Regra de Integração*. A determinação dos pesos e dos nós é feita de acordo com várias filosofias que se agrupam em duas subdivisões:

Fixas: A escolha de nós não depende do comportamento específico da função a ser integrada, mas apenas da regra a ser utilizada.

Adaptativa: A escolha dos pontos $\{x_j\}$ depende do comportamento da função, de modo que a densidade seja maior onde a função $f(x)$ varia com menos "suavidade".

Tanto na filosofia fixa como na adaptativa, empregamos vários tipos de regras. As mais importantes são:

Fórmulas de Newton-Cotes: Determinamos os pontos $x_j = x_0 + jh$ que são igualmente espaçados de uma distância h. Os pesos w_j são obtidos a partir de um polinômio de grau m que interpola f nos pontos $(x_j, f(x_j))$. Portanto, a regra obtida é exata para qualquer polinômio de grau menor ou igual a m.

Fórmulas de Gauss: Determinamos os pontos x_j e os pesos w_j de modo que a regra seja exata para qualquer polinômio de grau $p = 2n + 1$, onde n é o número de pontos a serem tomados no intervalo $[a, b]$. Os pontos x_j assim obtidos não são igualmente espaçados.

Fórmulas baseadas nos métodos de extrapolação do limite: As fórmulas Newtonianas podem apresentar uma convergência lenta. Uma forma de se aumentar a velocidade de convergência é aplicar uma fórmula de Newton-Cotes para $h = h_j$, $h_{j+1} < h_j$, obtendo-se uma sequência de aproximações da integral $\int_a^b f(x)dx$.

As integrais a serem calculadas podem ser próprias ou impróprias, convergentes ou não. As integrais impróprias são aquelas nas quais o intervalo de integração ou integrando são ilimitados. Tais integrais são definidas como um limite de integrais próprias, como está a seguir:

i. $\int_a^\infty f(x)dx = \lim_{x\to\infty} \int_a^x f(x)dx$, quando o limite existe

ii. $\int_{-\infty}^b f(x)dx = \lim_{x\to-\infty} \int_x^b f(x)dx$

iii. No caso de integrando não limitado, f é definida no intervalo (a, b) que é ilimitada numa vizinhança de a, então: $\int_a^b f(x)dx = \lim_{r\to a^+} \int_r^b f(x)dx$, quando o limite existe.

Além dos problemas anteriores, podemos ter integrais próprias, convergentes, porém mal comportadas. Isto ocorre quando a função não tem um comportamento polinomial, apresenta picos ou oscilações frequentes. A Figura 7.1 ilustra as questões relativas a funções própias e impróprias.

7.3 Fórmulas Newtonianas

7.3.1 Considerações Iniciais

As fórmulas Newtonianas são de aplicação mais simples quando temos a expressão de f ou quando obtemos uma tabela de pontos dados experimentalmente. As fórmulas dadas pela interpolação de f por polinômios de grau 1, 2 ou m podem ser aplicadas no intervalo $[a, b]$ constituindo regras simples, ou em subdivisões $[x_j, x_{j+1}]$ do intervalo $[a, b]$ formando regras compostas.

As fórmulas Newtonianas podem ser:

Fechadas: Quando o integrando f é calculado em $x_0 = a$ e $x_m = b$ sendo que a função f deve ser definida nestes pontos.

Abertas: Quando o integrando não é avaliado em ambas as extremidades do intervalo $[a, b]$ e sim em pontos próximos, assim $x_{m-r} = a$ e $x_{m+r} = b$ e $0 < r \leqslant m$ são utilizados quando há descontinuidade nos extremos.

Com termos de correção: O integrando é avaliado em pontos x_j fora do intervalo $[a, b]$ para fornecer uma correção ao valor calculado por uma regra fechada.

7.3.2 Regra dos Retângulos

Seja o intervalo finito $[a, b]$ no eixo x, que é particionado em n subintervalos $[x_j, x_{j+1}]$, $j = 1, \ldots, n$, onde $x_1 = a$, $x_{n+1} = b$, e $h_j = x_{j+1} - x_j$. Seja f uma função contínua, cuja integral não é conhecida. Nosso objetivo é calcular $F(x) = \int_a^b f(x)dx$ pelo cálculo das áreas de retângulos. Este procedimento é ilustrado na Figura 7.2, a qual exemplifica três tipos de regras. Na Figura 7.2(a), a área de cada retângulo é dada por $A = f(x_j)h_j$, na Figura 7.2(b) a área é dada por $A = f(x_{j+1})h_j$, e por fim na Figura 7.2(c) a área é dada por $A = f(\frac{x_j + x_{j+1}}{2})h_j$. Em qualquer das escolhas, a soma das áreas dos retângulos será uma aproximação de $\int_a^b f(x)dx$ que denotaremos por $I(f)$:

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

Considerando um intervalo de intregração $[a, b]$ subdividido em n subintervalos, teremos:

$$\begin{aligned} I(f) &\cong \sum_{j=1}^{n} I_j \\ I_j &\cong \int_{x_j}^{x_{j+1}} f(x)dx,\ j = 1, \ldots, n \end{aligned}$$

onde I_j é área do j-ésimo retângulo, sendo dada por uma das três fórmulas acima.

Duas regras para integração são:

Regra Simples: Uma fórmula simples para aproximação de $I(f)$ é utilizar apenas um retângulo, o que resulta nas expressões abaixo dependendo de como o retângulo é obtido:

$$\begin{aligned} I(f) &\cong f(a)(b-a) \\ I(f) &\cong f(b)(b-a) \\ I(f) &\cong f(\frac{a+b}{2})(b-a) \end{aligned}$$

Regra Composta: O intervalo $[a,b]$ é subdividido em n sub-intervalos. Pela regra dos retângulos, a integral será indicada por $R(h_j)$ e as regras de integração são:

$$\begin{aligned} R(h_j) &= \sum_{j=1}^{n} f(x_j)h_j \\ R(h_j) &= \sum_{j=1}^{n} f(x_{j+1})h_j \\ R(h_j) &= \sum_{j=1}^{n} f(\frac{x_j + x_{j+1}}{2})h_j \end{aligned}$$

variando conforme o tipo de retângulo, onde $x_{j+1} = x_j + h_j$. No caso em que $h_j = h$ é uma constante, então temos:

$$\begin{aligned} R(h_j) &= h\sum_{j=1}^{n} f(x_j) \\ R(h_j) &= h\sum_{j=1}^{n} f(x_{j+1}) \\ R(h_j) &= h\sum_{j=1}^{n} f(\frac{x_j + x_{j+1}}{2}) \end{aligned}$$

e consequentemente $x_{j+1} = x_j + h$.

7.3.3 Regra dos Trapézios

Se aproximarmos f por um polinômio $f^*(x)$ de grau 1 ao invés de um polinômio de grau zero, como foi realizado na regra dos retângulos, teremos:

$$\begin{aligned} f^*(x) &= \left[\frac{f(a)-f(b)}{a-b}\right]x + \frac{af(b)-bf(a)}{a-b} \\ &= \frac{f(a)x-bf(a)}{a-b} + \frac{af(b)-f(b)x}{a-b} \\ &= f(a)\frac{x-b}{a-b} + f(b)\frac{a-x}{a-b} \end{aligned}$$

que pode ser colocado na forma:

$$f^*(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}$$

Utilizando a aproximação linear $f^*(x)$ de $f(x)$, podemos verificar que:

$$\begin{aligned} \int_a^b f(x)dx &\cong \int_a^b f^*(x)dx \\ &= \int_a^b f(a)\frac{x-b}{a-b}dx + \int_a^b f(b)\frac{x-a}{b-a}dx \\ &= \frac{f(a)}{a-b}\int_a^b (x-b)dx + \frac{f(b)}{b-a}\int_a^b (x-a)dx \\ &= \frac{f(a)}{a-b}\left[\frac{(x-b)^2}{2}\right]_a^b + \frac{f(b)}{b-a}\left[\frac{(x-a)^2}{2}\right]_a^b \\ &= \frac{f(a)}{a-b}\left[-\frac{(a-b)^2}{2}\right] + \frac{f(b)}{b-a}\frac{(b-a)^2}{2} \\ &= \frac{-f(a)(a-b)}{2} + \frac{f(b)(b-a)}{2} \\ &= \frac{-f(a)(a-b)+f(b)(b-a)}{2} \\ &= \frac{b-a}{2}[f(a)+f(b)] \end{aligned}$$

Ou seja:

$$\int_a^b f(x)dx \cong \frac{b-a}{2}[f(a)+f(b)]$$

que corresponde à *regra simples do trapézio*, conforme ilustração na Figura 7.3.

Se subdividirmos o intervalo $[a, b]$ em n subintervalos e em cada um deles aproximarmos f por uma reta teremos a regra dos trapézios composta. Indicando por $T(h_j)$ a aproximação de $I(f)$ pela regra composta dos trapézios, teremos:

$$\begin{aligned} T(h_j) &= \sum_{j=1}^{n} T_j(h_j) \\ &= \sum_{j=1}^{n} \frac{f(x_j) + f(x_{j+1})}{2} h_j \end{aligned}$$

onde $h_j = x_{j+1} - x_j$, $j = 1, \ldots, n$. Se $h_j = h$, para todo j, podemos simplificar a expressão, obtendo:

$$T(h) = h\left[\frac{f(x_1) + f(x_2)}{2} + \frac{f(x_2) + f(x_3)}{2} + \ldots + \frac{f(x_n) + f(x_{n+1})}{2}\right]$$

ou ainda,

$$T(h) = \frac{h}{2}[f(x_1) + 2f(x_2) + 2f(x_3) + \ldots + 2f(x_n) + f(x_{n+1})]$$

7.3.4 Regra de Simpson

Se aproximarmos f por um polinômio f^* de grau 2 (uma parábola) teremos a chamada *Regra de Simpson*. Porém, para interpolarmos f por uma parábola precisamos de 3 pontos para construirmos a fórmula da regra simples. Sejam a e b dois pontos dados e y_m, o ponto médio dado por $y_m = \frac{a+b}{2}$. Pelo polinômio de Lagrange temos que:

$$\begin{aligned} f(x) &\cong f^*(x) \\ &= f(a)\frac{(x-b)(x-y_m)}{(a-b)(a-y_m)} + f(y_m)\frac{(x-a)(x-b)}{(y_m-a)(y_m-b)} + f(b)\frac{(x-a)(x-y_m)}{(b-a)(b-y_m)} \end{aligned}$$

Podemos obter $f^*(x)$ através do polinômio de Gregory-Newton, usando as diferenças finitas:

$$f^*(x) = f(a) + (x-a)\frac{\Delta f(a)}{h} + (x-a)(x-y_m)\frac{\Delta^2 f(a)}{2h^2}$$

onde,

$$\begin{aligned}\Delta f(a) &= f(y_m) - f(a)\\ \Delta^2 f(a) &= f(b) - 2f(y_m) + f(a)\end{aligned}$$

Fazendo uma mudança de variável:

$$\begin{aligned}x(\alpha) &= a + \alpha h, \alpha \in [0,2]\\ \frac{dx}{d\alpha} &= h \Rightarrow dx = hd\alpha\\ x - a &= a + \alpha h - a \Rightarrow x - a = \alpha h\end{aligned}$$

De acordo com esta mudança de variável, temos que $x(0) = a$, $x(1) = y_m$, e $x(2) = b$ para $h = \frac{b-a}{2}$. Daí deduzimos que:

$$\begin{aligned}(x-a)(x-y_m) &= \alpha h\left[a + \alpha h - \frac{(a+a+2h)}{2}\right]\\ &= \alpha h\left[\frac{2a + 2\alpha h - 2a - 2h}{2}\right]\\ &= \alpha h\left[\alpha - 1\right]h = \alpha(\alpha-1)h^2\end{aligned}$$

Podemos então obter a integral aproximada:

$$\begin{aligned}\int_a^b f^*(x)dx &= \int_0^2 \left[f(a) + \alpha h\frac{\Delta f(a)}{h} + \frac{\alpha(\alpha-1)h^2\Delta^2 f(a)}{2h^2}\right] hd\alpha\\ &= \int_0^2 \left[f(a) + \alpha\Delta f(a) + \frac{\alpha(\alpha-1)\Delta^2 f(a)}{2}\right] hd\alpha\\ &= h\left\{[f(a)\alpha]_0^2 + \Delta f(a)\left[\frac{\alpha^2}{2}\right]_0^2 + \frac{\Delta^2 f(a)}{2}\left[\frac{\alpha^3}{3} - \frac{\alpha^2}{2}\right]_0^2\right\}\\ &= h[2f(a) + 2(f(y_m) - f(a)) + \frac{1}{3}(f(b) - 2f(y_m) + f(a))]\end{aligned}$$

Portanto

$$\int_a^b f^*(x)dx = \frac{h}{3}[f(a) + 4f(y_m) + f(b)],$$

onde: $y_m = \frac{a+b}{2}$ e $h = \frac{b-a}{2}$.

A aproximação quadrática $f^*(x)$ de $f(x)$ no intervalo $[a,b]$, com ponto médio em y_m, é ilustrada na Figura 7.5.

Regra Composta de Simpson

Seguindo a mesma técnica de integração da regra composta dos trapézios, mas desta fez utilizando o integrador de Simpson, obtemos a Regra Composta de Simpson:

$$\begin{aligned} S(h_j) &= \sum_{j=1}^{n} S_j(h) \\ &= \sum_{j=1}^{n} \frac{h_j}{3}[f(x_j) + 4f(y_j) + f(x_{j+1})] \end{aligned}$$

onde $y_j = \frac{x_j + x_{j+1}}{2}$ e $h_j = \frac{x_{j+1} - x_j}{2}$, $j = 1, \ldots, n$.

Expandindo a expressão de Simpson e assumindo que $h_j = h$ para todo j, podemos expressá-la na forma:

$$\begin{aligned} S(h) &= \frac{h}{3}\left[f(x_1) + 4f(y_1) + 2f(x_2) + 4f(y_2) + 2f(x_3) + \ldots \right. \\ &\qquad \left. + 2f(x_n) + 4f(y_n) + f(x_{n+1})\right]. \end{aligned}$$

7.3.5 Fórmula Geral das Regras Newtonianas

Podemos generalizar os procedimentos anteriores e aproximar f por um polinômio de grau m. Lembre que na regra dos retângulos utilizamos um polinômio interpolador de grau 0, na regra dos trapézios um polinômio de grau 1, e por último um polinômio de grau 2 na regra de Simpson. Ao adotarmos um polinômio de grau m, precisamos determinar $m+1$ pontos no intervalo $[a, b]$ para a aplicação da regra simples. Seja:

- $h > 0$ a distância entre os nós;
- $x_0 = a$ o nó inicial;
- $x_k = x_0 + hk$, $k = 0, \ldots, m$, os demais nós; e
- $f_k = f(x_k)$ o valor da função nos diferentes nós.

Com base nestas definições, a fórmula de interpolação de Newton nos dá:

$$\begin{aligned} f(x) &= \sum_{k=0}^{m} \left(\begin{array}{c} u \\ k \end{array} \right) \Delta^k f_0 + R_{m+1}, \text{ onde} \\ u &= \frac{x - x_0}{h} \\ \left(\begin{array}{c} u \\ k \end{array} \right) &= \frac{u(u-1)(u-2)\ldots(u-k+1)}{k!} \\ R_{m+1} &= \frac{h}{(m+1)!} u(u-1)\ldots(u-m) f^{m+1}(\eta), x_0 < \eta < x_m \end{aligned}$$

Integrando f e trocando a integral $\int$ com a somatória $\sum$, temos:

$$\begin{aligned} \int_a^b f(x)dx &\cong h \sum_{k=0}^{m} a_k \Delta^k f_0 + R_{m+1}, \text{ onde} \\ a_k &= \int_\alpha^\beta \left(\begin{array}{c} u \\ k \end{array} \right) du; \ \alpha = \frac{a - x_0}{h} \text{ e } \beta = \frac{b - x_0}{h} \\ R_{m+1} &= h^{m+1} \int_\alpha^\beta \left(\begin{array}{c} u \\ m+1 \end{array} \right) f^{(m+1)}(\eta(u))du \end{aligned}$$

7.3.6 Exemplo 1

Tomemos como tarefa o cálculo de $\int_0^1 e^{-x^2} dx$, com $n = 4$ e $n = 8$ através da Regra dos Trapézios.

Caso i, $n = 4$**:** Nesta situação os parâmetros e nós são como segue:

$$\begin{aligned} h &= \frac{b-a}{4} = 0.25 \\ x_1 &= 0 \\ x_2 &= 0.25 \\ x_3 &= 0.5 \\ x_4 &= 0.75 \\ x_5 &= 1.0 \end{aligned}$$

Usando a expressão:

$$T(h) = \frac{h}{2}[f(x_1) + 2f(x_2) + 2f(x_3) + \ldots + 2f(x_n) + f(x_{n+1})]$$

e substituindo os valores acima, obtemos:

$$\begin{aligned} T(h) &= 0.125[1 + 2 \times 0.9394130632 + 2 \times 0.778800783 \\ &\quad +2 \times 0.589788 + 0.367879441] \\ &= 0.742984098 \end{aligned}$$

Caso ii, $n = 8$**:** Nesta situação os parâmetros e nós são como segue:

$$\begin{aligned} h &= 0.125 \\ x_1 &= 0 \\ x_2 &= 0.125 \\ x_3 &= 0.25 \\ x_4 &= 0.375 \\ x_5 &= 0.5 \\ x_6 &= 0.625 \\ x_7 &= 0.75 \\ x_8 &= 0.875 \\ x_9 &= 1.0 \end{aligned}$$

Usando a expressão:

$$T(h) = \frac{h}{2}[f(x_1) + 2f(x_2) + 2f(x_3) + \ldots + 2f(x_n) + f(x_{n+1})]$$

obtemos:

$$\begin{aligned} T(0.125) &= \frac{0.125}{2}[f(x_1) + 2f(x_2) + 2f(x_3) + \ldots + 2f(x_8) + f(x_9)] \\ &= 0.745865615 \end{aligned}$$

Comparando os dois resultados, vimos que podemos confiar em dois dígitos de cada resultado, então:

$$\int_0^1 e^{-x^2} \cong 0.74$$

7.3.7 Exemplo 2

Calcular $f(x)$ sendo f a função tabelada a seguir, usando a regra de Simpson:

x_j	1.9	2.0	2.1	2.2	2.3
f_j	3.41773	3.76220	4.14431	4.56791	5.03722

Usando a regra de Simpson, temos:

$$\begin{aligned}\int_a^b f(x)dx &= \frac{h}{3}[f(x_1) + 4f(x_2) + 2f(x_3) + 4f(x_2) + \ldots + 4f(x_n) + f(x_{n+1})] \\ h &= 0.1 \\ n &= 4\end{aligned}$$

Substituindo os valores acima, obtemos:

$$\begin{aligned}S(f) &= \frac{0.1}{3}[3.41773 + 4 \times 3.76220 + 2 \times 4.14431 + 4 \times 4.56791 + 5.03722] \\ &= 1.68880\end{aligned}$$

7.3.8 Exemplo 3

Calcular $\int_1^2 x \ln x dx$, usando a regra de Simpson, para $n = 1$ e $n = 2$.

Caso i, $n = 1$**:** Nesta situação, os parâmetros são conforme segue:

$$\begin{aligned}h &= \frac{1}{2} \\ S(1) &= \frac{h}{3}[f(1) + 4f(1.5) + f(2)] \\ &= \frac{1}{6}[0 + 2.432790649 + 1.386294361] \\ &= 0.636514163\end{aligned}$$

Caso ii, $n = 2$**:** Nesta situação, os parâmetros são conforme segue:

$$\begin{aligned}h &= \frac{1}{4} \\ S(0.25) &= \frac{h}{3}[f(1) + 4f(1.25) + 2f(1.5) + 4f(1.75) + f(2)] \\ &= \frac{0.25}{3}[0 + 1.115717756 + 1.216395324 + 3.917310515 + 1.386294361] \\ &= 0.636309829\end{aligned}$$

7.4 Estimativas de Erros

Para transformar a expressão abaixo numa igualdade:

$$\int_a^b f(x)dx \cong \sum_{j=1}^{n+1} w_j f(x_j)$$

consideraremos o erro que estamos cometendo. Embora o erro não possa ser calculado exatamente, em muitos casos ele pode ser estimado com boa precisão. O processo de integração numérica constitui um problema bem condicionado em princípio. É claro que, ao aproximarmos f por um polinômio f^*, estamos cometendo um erro mas se observarmos a Figura 7.6 veremos que a soma dos erros se anula à medida que n aumenta.

Adotaremos a notação abaixo para erros:

- E_{TTS} indicará o erro de truncamento da regra dos trapézios simples; e
- E_{TTC} indicará o erro de truncamento da regra dos trapézios composta.

7.4.1 Erro de Truncamento na Regra dos Trapézios Simples

Levando os erros em consideração, a integral pode ser colocada na forma:

$$\int_a^b f(x)dx = \frac{(b-a)}{2}[f(a)+f(b)] + E_{TTS}$$

Teorema 7.1 *Se $f(x)$ é duas vezes diferenciável em $[a,b]$, então o erro de truncamento E_{TTS} é dado por:*

$$E_{TTS} = -\frac{h^3}{12}f''(\xi), \text{ onde } \xi \in [a,b]$$

Exemplo de Aplicação

Calcular a integral $I = \int_1^2 \frac{e^{-x}}{x}dx$ pela regra dos trapézios simples.

$$\begin{aligned} I &= T(1) \\ &= \frac{1}{2}[f(1)+f(2)] \\ &= 0.5(3.678794412 \times 10^{-1} + 6.76676416 \times 10^{-2}) \\ &= 0.5(4.355470828 \times 10^{-1}) \\ &= 2.1777735414 \times 10^{-1} \end{aligned}$$

O valor exato para 12 casas decimais é $2.170483423687x10^{-1}$ e, portanto, o erro absoluto é:

$$|2.1777735414 \times 10^{-1} - 2.170483423687 \times 10^{-1}| = 4.729 \times 10^{-2}$$

Comparando o erro absoluto com o erro indicado pelo Teorema 7.1, precisamos inicialmente calcular as derivadas:

$$\begin{aligned} f'(x) &= -\frac{e^{-x}}{x^2} - \frac{e^{-x}}{x} \\ &= -\left(\frac{1}{x} + \frac{1}{x^2}\right) e^{-x} \\ f''(x) &= \frac{e^{-x}}{x} + \frac{e^{-x}}{x^2} + 2\frac{e^{-x}}{x^3} + \frac{e^{-x}}{x^2} \\ &= \left(\frac{1}{x} + \frac{2}{x^2} + \frac{2}{x^3}\right) e^{-x} \end{aligned}$$

Fazendo $\xi = 1$, $\xi \in [1, 2]$ temos que:

$$f''(\xi) = 5e^{-1} = 1.839$$

logo, o erro de truncamento previsto pelo teorema é $|\frac{1^3}{12}.f''(\xi)| = 0.15325$. Portanto, confirma-se que o erro absoluto é menor que o previsto.

7.4.2 Erro de Truncamento na Regra dos Trapézios Composta

Teorema 7.2 *Se f é duas vezes continuamente diferenciável em $[a, b]$, então o erro de truncamento da fórmula composta dos trapézios, para n subintervalos, é dado por:*

$$E_{TTC} = -\frac{h^2}{12}(b-a)f''(\xi),\ \xi \in [a, b]$$

Prova. Seja agora $h = \frac{b-a}{n}$ e $n = \frac{b-a}{h}$. Para cada subintervalo $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$, temos:

$$\begin{aligned} \int_{x_{i-1}}^{x_i} f(x)dx &= \frac{h}{2}[f(x_i) + f(x_{i-1})] + E_{TTi},\ \text{onde} \\ E_{TTi} &= -\frac{h^3}{12}f''(\xi_i),\ \xi_i \in [x_{i-1}, x_i] \end{aligned}$$

Uma vez que:

$$\int_a^b f(x)dx = \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)dx$$

e pela regra composta, temos que:

$$\int_a^b f(x)dx = \frac{h}{2}[f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)] + \sum_{i=1}^{n} E_{TTi}$$

mas,

$$\begin{aligned}
\sum_{i=1}^{n} E_{TTi} &= \sum_{i=1}^{n} -\frac{h^3}{12} f''(\xi_i),\ \xi_i \in [x_{i-1}, x_i] \\
&= -\frac{h^3}{12} \sum_{i=1}^{n} f''(\xi_i) \\
&= -\frac{(b-a)^3}{12n^3} \sum_{i=1}^{n} f''(\xi_i)
\end{aligned}$$

Como $f''(x)$ é contínua, $f''(x)$ assume todos os valores entre seus máximos e mínimos em $[a, b]$. Portanto, existe algum $\xi \in [a, b]$ tal que:

$$f''(\xi) = \frac{\sum_{i=1}^{n} f''(\xi_i)}{n}$$

Logo,

$$\begin{aligned}
E_{TTC} &= -\frac{(b-a)^3}{12n^3} n f''(\xi) \\
&= -\frac{(b-a)^3}{12n^2} f''(\xi) \\
&= -\frac{h^2}{12}(b-a) f''(\xi)
\end{aligned}$$

■

Exemplo de Aplicação

Considerando o exemplo anterior, vamos calcular $\int_1^2 \frac{e^{-x}}{x} dx$, com $n = 2, 4, \ldots, 256$. Os resultados destes cálculos, juntamente com o erro absoluto e os limites de erro calculados são listados na tabela abaixo.

h	n	Valor calculado	Erro absoluto	Limite de erros
1	1	0.2177735413	4.729×10^{-2}	1.53×10^{-1}
5.0×10^{-1}	2	0.1832634907	1.280×10^{-2}	3.83×10^{-2}
2.5×10^{-1}	4	0.1737575538	3.270×10^{-3}	9.58×10^{-3}
1.25×10^{-1}	8	0.1715074075	8.240×10^{-4}	2.40×10^{-3}
6.25×10^{-2}	16	0.1706897700	2.060×10^{-4}	5.60×10^{-4}
$\vdots$				
3.906×10^{-3}	256	0.1704847700	8.070×10^{-7}	2.34×10^{-6}

Podemos observar que cada vez que o número de intervalos n é dobrado, o erro absoluto é reduzido por um fator de aproximadamente 4, o que está de acordo com o resultado do teorema.

7.4.3 Estimação Numérica do Erro de Truncamento da Regra dos Trapézios

O que fazer quando $f''(x)$ não estiver disponível? Duas possibilidades são:

i. Calcular $f''(x)$ numericamente

ii. Calcular $T(h)$ e $T(\frac{h}{2})$ e comparar os resultados

No caso (i), a segunda derivada $f''(x)$ é calculada numericamente pela série de Taylor se f é suficientemente diferenciável:

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + O(h^2)$$

Logo, um limite de $|f''(x)|$ pode ser calculado por:

$$\max_{1 \leqslant j \leqslant n} \frac{f(x_{j+1}) - 2f(x_j) + f(x_{j-1})}{h^2}$$

O limite acima pode ser útil no caso da integração para pontos tabelados igualmente espaçados de h.

No caso (ii), podemos utilizar o Teorema 7.2:

$$\begin{aligned} I - T(h) &= -\frac{(b-a)}{12} h^2 f''(\xi_1),\ \xi_1 \in [a,b] \\ I - T(\frac{h}{2}) &= -\frac{(b-a)}{12} \frac{h^2}{4} f''(\xi_2),\ \xi_2 \in [a,b] \end{aligned}$$

Assumindo que $f''(\xi_1) = f''(\xi_2)$, temos que:

$$\begin{aligned} 4[I - T(\frac{h}{2})] &\cong I - T(h) \\ 4I - 4T(\frac{h}{2}) &\cong I - T(h) \\ 3I - 3T(\frac{h}{2}) &\cong T(\frac{h}{2}) - T(h) \\ I - T(\frac{h}{2}) &\cong \frac{T(\frac{h}{2}) - T(h)}{3} \end{aligned}$$

Chegamos à conclusão de que o erro de truncamento, ao calcularmos $T(\frac{h}{2})$ é, aproximadamente, a terça parte entre as duas aproximações $T(\frac{h}{2})$ e $T(h)$. Este método é particularmente vantajoso, pois, ao calcularmos $T(\frac{h}{2})$ podemos reutilizar os valores de f usados para calcular $T(h)$. Ou seja,

$$\begin{aligned} T(h) &= h\left[\frac{f(a)}{2} + f(a+h) + f(a+2h) + \ldots + f(a+(n-1)h) + \frac{f(b)}{2}\right] \\ T(\frac{h}{2}) &= \frac{h}{2}\left[\frac{1}{2}f(a) + f(a+\frac{h}{2}) + f(a+h) + f(a+\frac{3h}{2}) + f(a+2h)\right. \\ &\quad \left. + \ldots + f(a+(n-1)h) + f(a+(n-\frac{1}{2})h) + \frac{f(b)}{2}\right] \end{aligned}$$

Portanto,

$$T(\frac{h}{2}) = \frac{1}{2}T(h) + \frac{h}{2}\sum_{j=1}^{n} f(a + (j - \frac{1}{2})h)$$

logo o número de avaliações é reduzido pela metade.

7.5 Quadratura Gaussiana

Os métodos de integração numérica apresentados acima (a saber, regra dos retângulos, dos trapézios e de Simpson) tomam como base uma regra simples para escolha dos pontos de avaliação da função $f(x)$, onde $x_{j+1} = x_j + h$. Esses métodos são particularmente adequados para dados tabulados de forma regular, tais como medidas de laboratório e valores obtidos de programas de computador que produzem tabelas.

Se, por outro lado, tivermos a liberdade de escolher os pontos nos quais a função $f(x)$ é avaliada, então uma escolha cuidadosa pode levar a uma maior precisão da avaliação da função. Este método, conhecido por *Integração*

Gaussiana ou *Integração de Gauss-Legendre*, apresenta outras vantagens em várias situações. Na avaliação da integral:

$$\int_a^b f(x)dx \tag{7.2}$$

não é necessário avaliar a função nos pontos extremos do intervalo, a e b. Esta propriedade é útil quando se avalia integrais impróprias, tais como aquelas com limites infinitos.

Para efeitos de simplificação, fazemos uma mudança de variável com

$$x = a + \frac{b-a}{2}(t+1)$$

o que implica:

$$dx = \frac{b-a}{2}dt$$

Além disso, para $x = a$ temos $t = -1$ e para $x = b$ temos $t = 1$. Por meio da mudança de variável e normalização do intervalo de integração, podemos representar (7.2) por:

$$\frac{(b-a)}{2}\int_{-1}^{+1} F(t)dt \tag{7.3}$$

onde:

$$F(t) = f(a + \frac{b-a}{2}(t+1)) \tag{7.4}$$

Logo, daqui em diante, vamos considerar apenas o problema normalizado:

$$\int_{-1}^{+1} F(t)dt \tag{7.5}$$

A forma mais simples de integração Guassiana se alicerça na escolha de um polinômio aproximador ótimo do integrando $F(t)$ sobre o intervalo $[-1, +1]$. Os detalhes da determinação desse polinômio, ou seja, a determinação dos coeficientes de t no polinômio, serão abordados mais à frente. Faremos uma aproximação de (7.5) por valores da função e ponderações para valores de $t \in [-1, +1]$, conforme a expressão:

$$\int_{-1}^{+1} F(t)dt = \sum_{j=1}^{n} \omega_j F(t_j) + E_n^{(G)},\ n \geqslant 1 \tag{7.6}$$

onde ω_j e t_j são escolhidos de maneira que a regra seja exata para polinômios de grau $2n-1$. Regras desta natureza são ditas Gaussianas e os pontos t_j não são necessariamente igualmente espaçados. No que segue discutimos o cálculo das ponderações ω_j e dos pontos de avaliação t_j.

7.5.1 Regra de Gauss de Primeira Ordem

Nesta situação, $n = 1$ e aproximaremos (7.5) de forma exata com um polinômio de grau $p = 2n - 1 = 1$, o que leva à expressão:

$$\int_{-1}^{+1} F(t)dt = \omega_1 F(t_1) + E_1^{(G)} \tag{7.7}$$

Vamos encontrar os valores para situações onde $F(t) = t^k$ para $k \in \{0, 1\}$.

Para $k = 0$:

$$\int_{-1}^{+1} F(t)dt = \int_{-1}^{+1} 1dt = [t]_{-1}^{+1} = 2$$

O que leva a:

$$\omega_1 F(t_1) = 2 \;\Leftrightarrow\; \omega_1 t_1^0 \;\Leftrightarrow\; \omega_1 = 2$$

Para $k = 1$:

$$\int_{-1}^{+1} F(t)dt = \int_{-1}^{+1} tdt = [\frac{t^2}{2}]_{-1}^{+1} = 0$$

Daí segue que:

$$\omega_1 F(t_1) = 0 \;\Leftrightarrow\; \omega_1 t_1^1 = 0 \;\Leftrightarrow\; t_1 = 0$$

Dos desenvolvimentos acima, concluímos que:

$$\int_{-1}^{+1} F(t)dt = 2F(0) + E_1^{(G)} \tag{7.8}$$

7.5.2 Regra de Gauss de Segunda Ordem

Nesta situação, $n = 2$ e aproximaremos (7.5) com a expressão:

$$\int_{-1}^{+1} F(t)dt = \omega_1 F(t_1) + \omega_2 F(t_2) + E_2^{(G)} \tag{7.9}$$

Agora, ω_1, ω_2, t_1 e t_2 devem ser escolhidos de forma que (7.9) seja exata para polinômios de grau até $p = 2n - 1 = 3$. Vamos encontrar os valores para situações onde $F(t) = t^k$ para $k \in \{0, 1, 2, 3\}$.

Para $k = 0$:

$$\begin{aligned} \int_{-1}^{+1} F(t)dt &= \int_{-1}^{+1} t^0 dt = \int_{-1}^{+1} dt = 2 \\ &= \omega_1 F(t_1) + \omega_2 F(t_2) = \omega_1 t_1^0 + \omega_2 t_2^0 = \omega_1 + \omega_2 \end{aligned} \tag{7.10}$$

Para $k = 1$:

$$\begin{aligned}\int_{-1}^{+1} F(t)dt &= \int_{-1}^{+1} t^1 dt = [\frac{t^2}{2}]_{-1}^{+1} = 0 \\ &= \omega_1 F(t_1) + \omega_2 F(t_2) = \omega_1 t_1^1 + \omega_2 t_2^1 = \omega_1 t_1 + \omega_2 t_2 \end{aligned} \quad (7.11)$$

Para $k = 2$:

$$\int_{-1}^{+1} F(t)dt = \int_{-1}^{+1} t^2 dt = [\frac{t^3}{3}]_{-1}^{+1} = \frac{2}{3} = \omega_1 t_1^2 + \omega_2 t_2^2 \quad (7.12)$$

Para $k = 3$:

$$\int_{-1}^{+1} F(t)dt = \int_{-1}^{+1} t^3 dt = [\frac{t^4}{4}]_{-1}^{+1} = 0 = \omega_1 t_1^3 + \omega_2 t_2^3 \quad (7.13)$$

A partir das equações (7.10)–(7.13) chegamos a um sitema de equações não-lineares:

$$\omega_1 + \omega_2 = 2 \quad (7.14)$$

$$\omega_1 t_1 + \omega_2 t_2 = 0 \quad (7.15)$$

$$\omega_1 t_1^2 + \omega_2 t_2^2 = \frac{2}{3} \quad (7.16)$$

$$\omega_1 t_1^3 + \omega_2 t_2^3 = 0 \quad (7.17)$$

Podemos resolver (7.14)–(7.17) fazendo, primeiramente, $\omega_1 = \omega_2 = 1$, o que resolve (7.14). Daí, obtemos a partir de (7.15) que $t_1 = -t_2$. Substiuindo estes resultados em (7.16), deduzimos que $t_1^2 + t_2^2 = 2t_2^2 = 2/3 \Rightarrow t_2 = 1/\sqrt{3}$ e, por sua vez, descobrimos que $t_1 = -1/\sqrt{3}$. Observe ainda que os valores obtidos para os parâmetros satisfazem a equação (7.17). Com base nestes desenvolvimentos, concluímos que:

$$\begin{aligned}\int_{-1}^{+1} F(t)dt &= \omega_1 F(t_1) + \omega_2 F(t_2) + E_2^{(G)} \\ &= F(-\frac{1}{\sqrt{3}}) + F(+\frac{1}{\sqrt{3}}) + E_2^{(G)} \end{aligned} \quad (7.18)$$

7.5.3 Exemplo de Aplicação

Como exemplo, vamos calcular of valor da integral definida $\int_2^6 \frac{x^3}{3} dx$ por meio da quadratura Gaussiana de segunda ordem.

Etapa 1: A primeira etapa corresponde à mudança de variável. Com $a = 2$, $b = 6$ e $f(x) = x^3/3$, obtemos:

$$\begin{aligned} x &= a + \frac{(b-a)}{2}(t+1) \\ &= 2 + \frac{(6-2)}{2}(t+1) \\ &= 4 + 2t \\ &\Rightarrow dx = 2dt \\ F(t) &= f(a + \frac{b-a}{2}(t+1)) \\ &= f(4+2t) \end{aligned}$$

Chegamos assim à equivalência:

$$\int_2^6 \frac{x^3}{3}dx = \int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{+1} F(t)dt = 2\int_{-1}^{+1} \frac{(2t+4)^3}{3}dt$$

Etapa 2: Com $n = 2$, temos $\int_{-1}^{+1} F(t)dt \approx F(-\frac{1}{\sqrt{3}}) + F(+\frac{1}{\sqrt{3}})$. Calculando estes valores, obtemos:

$$\begin{aligned} F(-\frac{1}{\sqrt{3}}) &= \frac{(-2/\sqrt{3}+4)^3}{3} = 7.678358 \\ F(+\frac{1}{\sqrt{3}}) &= \frac{(2/\sqrt{3}+4)^3}{3} = 45.655075 \end{aligned}$$

Então, calculamos que:

$$\begin{aligned} \int_2^6 \frac{x^3}{3}dx = 2\int_{-1}^{+1} \frac{(2t+4)^3}{3}dt &= 2 \times (7.678358 + 45.655075) \\ &= 2 \times 53.333333 \\ &= 106.666666 \end{aligned}$$

Analiticamente,

$$\int_2^6 \frac{x^3}{3}dx = \left[\frac{x^4}{4\times 3}\right]_2^6 = \frac{6^4 - 2^4}{12} = 106.666666$$

verificando o erro nulo para o caso de um polinômio de grau 3.

Tabela 7.1: Ponderações e pontos de amostragem para quadratura Gaussiana

n	ω_j	t_j
1	$\omega_1 = 2$	$t_1 = 0$
2	$\omega_1 = \omega_2 = 1$	$t_1 = -1/\sqrt{3}$ e $t_2 = 1/\sqrt{3}$
3	$\omega_1 = \omega_2 = 5/9$ $\omega_3 = 8/9$	$t_1 = -\sqrt{0.6}$ e $t_2 = \sqrt{0.6}$ $t_3 = 0$
4	$\omega_1 = \omega_4 = 0.3478548451$ $\omega_2 = \omega_3 = 0.6521451549$	$t_1 = -0.8611363116$ e $t_4 = -t_1$ $t_2 = -0.3399810436$ e $t_3 = -t_2$

7.5.4 Quadratura de Ordem Superior

Através de desenvolvimento semelhante ao apresentado acima, é possível encontrar as ponderações ω_j e os pontos de avaliação t_j para a integração Gaussiana de ordem 3, 4, e maior. Abaixo apresentamos na Tabela 7.5.4 os valores já calculados destes parâmetros para quadratura de ordem até 4.

Com relação aos erros cometidos pela regra da quadratura Gaussiana, estimativas dos erros podem ser estabelecidas conforme a ordem da quadratura. Para $n = 2$,

$$E_2^{(G)} = \frac{1}{135} F^{(iv)}(\beta), \qquad \text{para } \beta \in [-1, +1]$$

Portanto,

$$|E_2^{(G)}| \leqslant \frac{1}{135} \max\{|F^{(iv)}(\beta)| : \beta \in [-1, +1]\}$$

Para o caso de $f(x)$ corresponder a um polinômio de grau igual ou inferior a 3, o erro produzido pela quadratura é nulo.

7.6 Referências

O conteúdo deste capítulo foi inspirado nos textos de Cláudio e Marins [1] e de Dyer e Ip [3].

7.7 Exercícios

Exercício 7.1 Calcule $\int_1^2 e^x dx$ usando i) a regra dos retângulos simples e ii) a regra dos trapézios simples.

Exercício 7.2 Considere a Tabela 7.2 com pontos de uma função desconhecida $f(x)$, mas contínua. Tarefas:

a) Calcule $\int_0^4 f(x)dx$ usando a Regra de Simpson

b) Calcule $\int_0^4 f(x)dx$ usando a Regra dos Trapézios

c) Calcule $\int_0^4 f(x)dx$ usando a Regra dos Retângulos com altura dada pela média dos valores da função nos extremos de cada subintervalo.

Tabela 7.2: Pontos da função $f(x)$

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$f(x)$	-4271	-2522	499	1795	4358	7187	10279	13633	17217

Exercício 7.3 Seja $f(x)$ uma função duas vezes diferenciável no intervalo $[a, b]$. Então o erro de truncamento produzido pela regra dos trapézios simples, E_{TTS}, obtido ao se calcular $\int_a^b f(x)dx$ é dado por

$$E_{TTS} = -\frac{h^3}{12} f''(\xi), \quad \textit{onde } \xi \in [a, b].$$

Calcule o erro máximo de E_{TTS} gerado ao aplicarmos a regra dos trapézios para calcular $\int_1^2 \frac{e^{-x}}{x^2} dx$.

Exercício 7.4 Calcule o erro máximo de E_{TTS} gerado ao aplicarmos a regra dos trapézios simples para calcular $\int_{\pi/2}^{\pi} [2x^4 - 3x^3 + 2x^2 - x + 1 - \sin(x - \pi/2)]dx$.

Exercício 7.5 Desejamos calcular $\int_{-4}^{4} f(x)dx$ da forma mais precisa possível. Não conhecemos a função $f(x)$, mas sabemos os valores de f em alguns pontos, conforme Tabela 7.3. Sabendo que $f(x)$ é um polinômio de grau 6, como que você calcularia a integral? Basta explicar com clareza como que você resolveria o problema. Não é necessário calcular o valor da integral.

Tabela 7.3: Pontos da função $f(x)$

x	-3	-2	-1	0	1	2	3
$f(x)$	1339	171	7	1	3	79	751

Exercício 7.6 Calcule aproximações para as integrais I_1 e I_2 por meio dos métodos:

i. regra dos retângulos simples (extremo à esquerda);

ii. regra dos retângulos composta (extremo à esquerda), com $n = 10$;

iii. regra dos trapézios simples;

iv. regra dos trapézios composta, com $n = 10$;

v. regra de Simpson simples;

vi. regra de Simpson composta, com $n = 10$;

vii. quadratura Gaussiana de 2^a ordem; e

viii. quadratura Gaussiana de 3^a ordem.

$$
\begin{aligned}
I_1 &= \int_0^3 (3x^3 + 2x^2 + 2x - 2)dx \\
I_2 &= \int_0^5 \frac{e^{-x^2}}{x+1} dx
\end{aligned}
$$

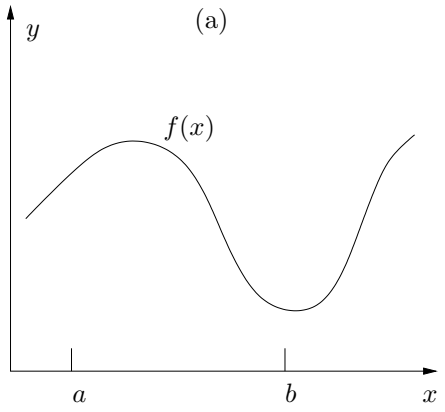
Outras tarefas:

i. calcule o valor analítico de I_1;

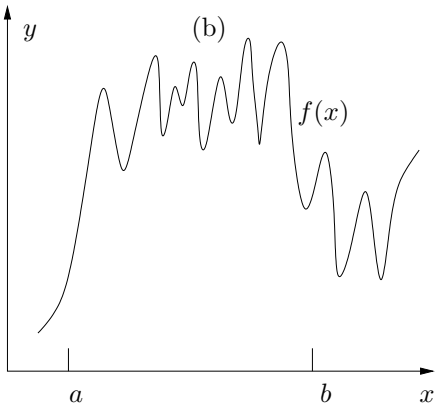
ii. calcule o erro absoluto das aproximações calculadas acima para I_1, itens (a)-(h) acima;

iii. calcule o erro máximo cometido com a regra dos trapézios simples com base nos resultados teóricos, assumindo que não se conhece o valor exato de I_1; e

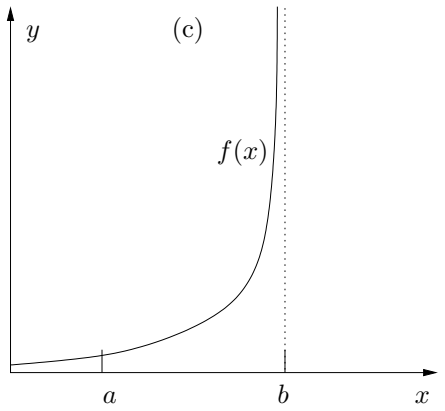
iv. calcule o erro máximo cometido com a regra dos trapézios composta com base nos resultados teóricos, assumindo que não se conhece o valor exato de I_1.



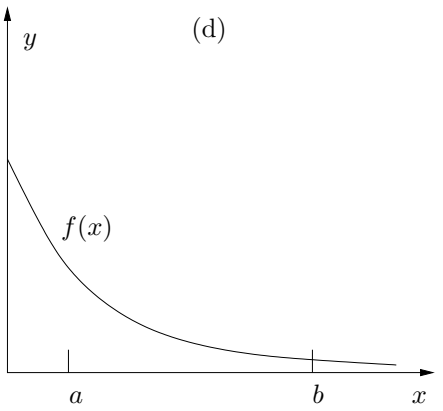
Integral própria
- Bem comportada
- Integrando suave, polinomial



Integral própria
- Mal comportada
  - Integrando não suave, com variações bruscas



Integral imprópria
- Intervalo de integração limitado
- Integrando não limitado



Integral imprópria
- Integrando pode ser limitado ou ilimitado

Figura 7.1: Comportamento de integrais

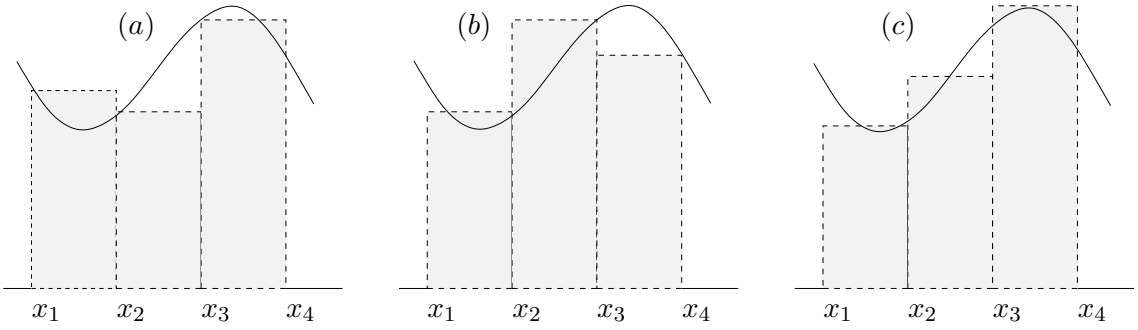


Figura 7.2: Regras dos retângulos

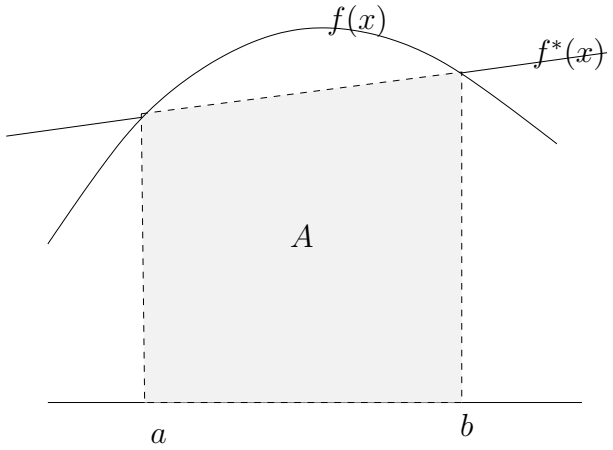


Figura 7.3: Regra simples do trapézio

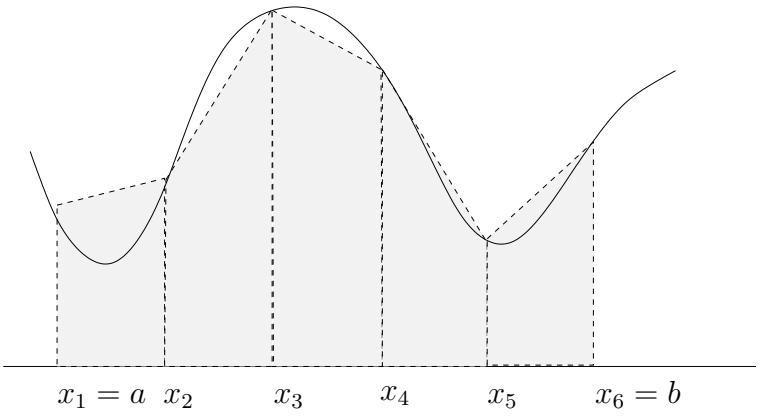


Figura 7.4: Regra composta do trapézio

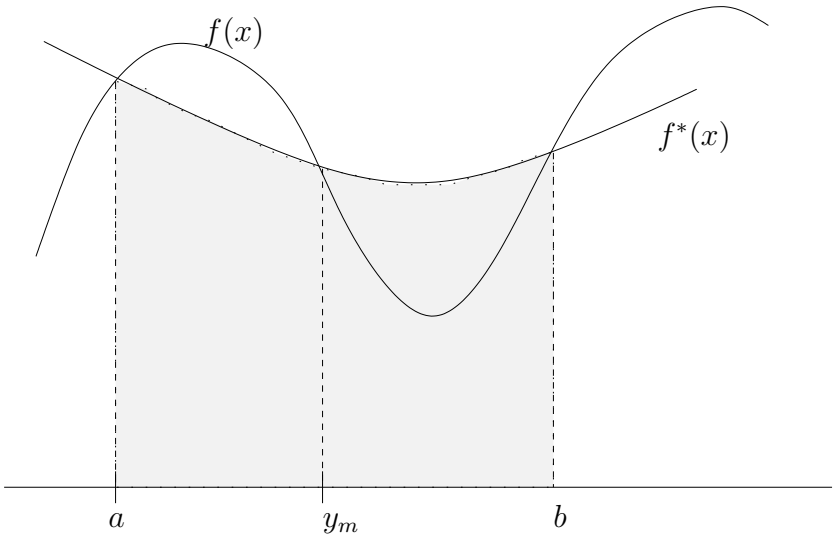


Figura 7.5: Regra de Simpsom

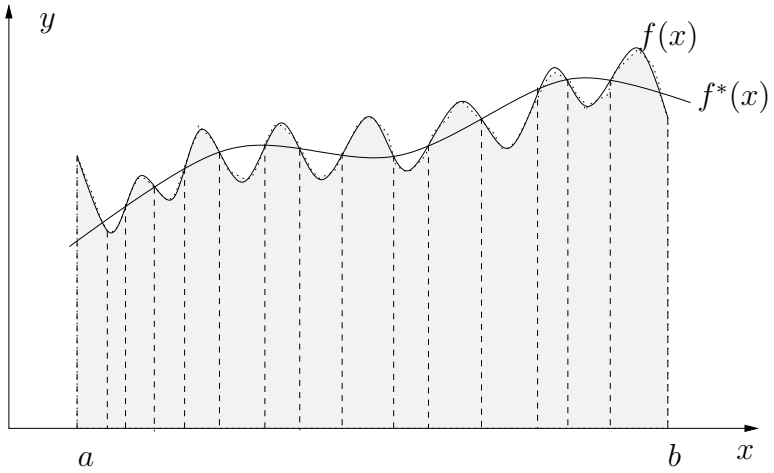


Figura 7.6: Cancelamento de erros

Capítulo 8

Resolução Numérica de Equações Diferenciais Ordinárias

Fenômenos físicos frequentemente envolvem relações entre uma variável independente x e uma variável dependente y, que não são fáceis ou mesmo possíveis de serem descritas como uma função da variável independente: $y = f(x)$. Por outro lado, podemos muitas vezes estabelecer a relação entre y e x através de seus valores e as derivadas da função desconhecida dy/dx. Em circuitos elétricos, por exemplo, desejamos encontrar a voltagem como uma função do tempo, $v(t)$, que pode ser escrita como uma relação das derivadas de v no tempo e as propriedades do circuito. Uma relação expressa como uma função da variável independente x, da variável dependente y e suas derivadas $y'(x)$, $y''(x)$, ... é dita *equação diferencial*. Uma relação que envolve derivadas até ordem n é dita *equação diferencial ordinária* (EDO), podendo ser colocada na forma matemática:

$$f(x, y(x), y'(x), \ldots, y^{(n)}(x)) = 0$$

Neste capítulo faremos uma breve introdução à modelagem de fenômenos físicos através de equações diferenciais, desenvolveremos ainda métodos para encontrar soluções numéricas para equações e sistemas de equações diferenciais ordinárias. No caso da variável independente ser o tempo t, o sistema de equações diferenciais ordinárias toma a forma:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \ldots, x_n) \\ \dot{x}_2 &= f_2(x_1, \ldots, x_n) \\ \vdots & \qquad \vdots \\ \dot{x}_n &= f_n(x_1, \ldots, x_n) \end{aligned}$$

Mais especificamente, estaremos interessados no problema de encontrar a trajetória $x(t)$, $t \in [0, T]$, a partir de um estado inicial $x(0) \in \mathbb{R}^n$ onde $x(t) = (x_1(t), \ldots, x_n(t))$.

8.1 Modelagem com Equações Diferenciais

Objetivando ilustrar a modelagem com equações diferencias, desenvolveremos a seguir modelos de sistemas dinâmicos.

8.1.1 Circuito RC

O circuito RC é composto de uma fonte de tensão, $v_i(t)$, em série com um resistor R e um capacitor C, conforme ilustração da Figura 8.1. Da Física, sabemos que a corrente no capacitor é proporcional à taxa de variação da tensão através do capacitor, matematicamente:

$$i(t) = C\frac{dv_c(t)}{dt}, \tag{8.1}$$

sendo a capacitância C a constante de proporcionalidade. Pela lei de Kirchoff, a soma das quedas dos potenciais ao longo da malha deve ser nulo, o que leva à expressão:

$$v_i(t) - Ri(t) - v_c(t) = 0 \tag{8.2}$$

Substituindo $i(t)$ em (8.2) pela relação (8.1), surge uma equação diferencial de primeira ordem:

$$v_i(t) - RC\frac{dv_c(t)}{dt} - v_c(t) = 0 \implies \frac{dv_c(t)}{dt} = -\frac{1}{RC}v_c(t) + \frac{1}{RC}v_i(t) \tag{8.3}$$

Considere o caso simples onde $v_i(t) = 0$ para todo t e $v_c(0) = v_c^o$, correspondendo à situação de descarga do capacitor. Então, a solução analítica de (8.3) pode ser obtida:

$$\begin{aligned} \frac{dv_c(t)}{dt} = -\frac{1}{RC}v_c(t) &\Leftrightarrow \frac{dv_c(t)}{v_c(t)} = -\frac{dt}{RC} \\ &\Leftrightarrow \int_{t=0}^{T} \frac{dv_c(t)}{v_c(t)} = \int_{t=0}^{T} -\frac{dt}{RC} \\ &\Leftrightarrow \ln v_c(t) = -\frac{t}{RC} + k \end{aligned} \tag{8.4}$$

onde k é uma constante. Portanto, a partir de (8.4), deduzimos que a tensão no capacitor decresce exponencialmente na taxa inversa de RC:

$$v_c(t) = e^{-\frac{t}{RC}+k} = e^k e^{-\frac{t}{RC}} = v_c^o e^{-\frac{t}{RC}} \tag{8.5}$$

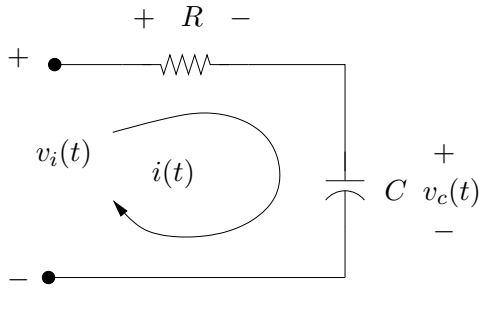


Figura 8.1: Circuito RC

Para um circuito RC onde $R = 2\ \Omega$, $C = 0.1\ F$ e $v_c(0) = 2\ V$, a curva de tensão no capacitor em função do tempo pode ser observada na Figura 8.2. Esta curva caracteriza a descarga da energia do capacitor que, por sua vez, é dissipada pelo resistor.

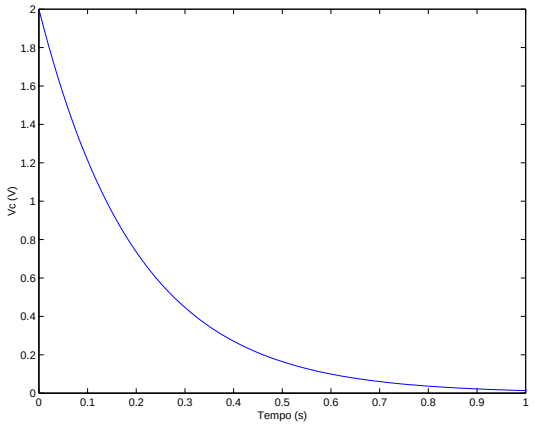


Figura 8.2: Curva de descarga do capacitor em um circuito RC

8.1.2 Circuito RLC

O circuito RLC consiste de uma fonte de tensão $v_i(t)$ em série com um resistor R, um indutor L e um capacitor C, de acordo com o diagrama da Figura 8.3. A lei de Kirchoff nos diz que a soma das quedas dos potenciais ao longo da malha deve ser nulo, portanto:

$$v_i(t) - Ri(t) - L\frac{di(t)}{dt} - v_c(t) = 0 \tag{8.6}$$

Lembramos que a queda de tensão no indutor é proporcional à taxa de variação da corrente, sendo L a constante de proporcionalidade. Uma vez que a

corrente através do capacitor é proporcional à taxa de variação da queda de tensão no capacitor, obtemos juntamente com (8.6), o sistema de equações diferenciais de 1^a ordem:

$$v_i(t) = Ri(t) + L\frac{di(t)}{dt} + v_c(t) \tag{8.7}$$

$$i(t) = C\frac{dv_c(t)}{dt} \tag{8.8}$$

que, por sua vez, pode ser colocado na forma matricial:

$$\begin{bmatrix} di(t)/dt \\ dv_c(t)/dt \end{bmatrix} = \begin{bmatrix} -R/L & -1/L \\ 1/C & 0 \end{bmatrix} \begin{bmatrix} i(t) \\ v_c(t) \end{bmatrix} + \begin{bmatrix} 1/L \\ 0 \end{bmatrix} v_i(t) \tag{8.9}$$

Alternativamente, o sistema de primeira ordem (8.9) pode ser colocado como uma equação diferencial de segunda ordem, bastando para isto substituir (8.8) em (8.7):

$$v_i(t) = LC\frac{d^2v_c(t)}{dt^2} + RC\frac{dv_c(t)}{dt} + v_c(t) \tag{8.10}$$

Aqui ilustramos como se transforma uma EDO de ordem n em um sistema EDO de primeira ordem com n equações. Definindo x como variável de estado:

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} v_c(t) \\ dv_c(t)/dt \end{bmatrix}$$

e estabelecendo $u(t)$ como a entrada e $y(t)$ como a saída, teremos:

$$u(t) = v_i(t), \quad y(t) = v_c(t)$$

Note que a entrada é a tensão $v_i(t)$, enquanto a saída (o que é observado) é a queda de tensão no capacitor. Agora podemos expressar a EDO (8.10) de 2^a ordem em um sistema EDO de 1^a ordem:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1/LC & -R/L \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1/LC \end{bmatrix} u(t) \tag{8.11}$$

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{8.12}$$

As equações diferenciais do circuito RLC, conforme (8.11)–(8.12), fazem parte dos sistemas de equações diferenciais lineares:

$$\dot{x} = Ax + Bu \tag{8.13}$$

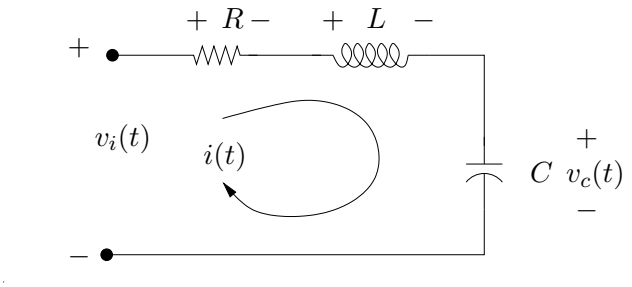


Figura 8.3: Circuito RLC

para $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times m}$ e $u \in \mathbb{R}^m$. Supondo que u é uma função de x como, por exemplo, $u = -Kx$ onde K é a matriz de ganhos, podemos assumir que (8.13) é da forma:

$$\dot{x} = Ax \tag{8.14}$$

Uma solução analítica para (8.14) pode ser obtida. Quando este sistema se reduz a uma equação, $\dot{x} = ax$, a solução é trivial, assumindo a forma $x(t) = e^{at}$. O caso geral não é muito diferente do caso particular, para tanto, definimos a função exponencial de matriz como:

$$e^{At} = I + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \ldots = \sum_{k=0}^{\infty} \frac{(At)^k}{k!} \tag{8.15}$$

Então, $x(t) = e^{At}x_0$ é a solução de (8.14) com $x(0) = x_0$. Basta verificar que:

$$\begin{aligned}
\frac{d}{dt}\left[e^{At}\right]x_0 &= \frac{d}{dt}\left[I + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \ldots\right]x_0 \\
&= \left[0 + A + A^2t + \frac{1}{2!}A^3t^2 + \frac{1}{3!}A^4t^3 + \ldots\right]x_0 \\
&= A\left[I + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \ldots\right]x_0 \\
&= Ae^{At}x_0
\end{aligned} \tag{8.16}$$

Uma propriedade fundamental de sistemas dinâmicos sob a ação de controladores é a estabilidade, que pode ser entendida como a convergência do estado $x(t)$ para um ponto de equilíbrio x^*. De maneira simplificada, dizemos que o sistema (8.14) é estável se:

$$\lim_{t\to\infty} x(t) = x^*$$

Sob quais condições o sistema caracterizado pela equação $\dot{x} = ax$ é estável? É fácil de verificar que a estabilidade é garantida a partir de qualquer ponto inicial $x_0 = x(0)$ quando $a < 0$, o que equivale a dizer que $\lim_{t\to\infty} e^{at}x_0 = 0$.

O que podemos dizer sobre a convergência de um sistema multivariável caracterizado pelo sistema EDO $\dot{x} = Ax$? Convergência pode ser garantida quando e^{At}, em outras palavras, quando a série $I+At+A^2t^2/2!+A^3t^3/3!+\ldots$ é convergente, o que ocorre quando todos os autovalores de A tem parte real negativa.

A Figura 8.4 ilustra a resposta do circuito RLC para uma entrada nula, $u(t) = 0$, com $R = 2\ \Omega$, $C = 0.1\ F$, $L = 0.4\ H$, $v_c(0) = 2\ V$ e $\dot{v}_c(0) = 10\ V/s$. São dadas as curvas $v_c(t)$ e $\dot{v}_c(t)$ para $t \in [0, 4]\ s$. Observe que o circuito RLC é convergente para a origem a partir do ponto inicial dado. Na verdade, o circuito é estável como pode ser verificado calculando os autovalores de A, a saber $-2 \mp 4j$, os quais tem parte real negativa garantindo convergência a partir de qualquer estado inicial.

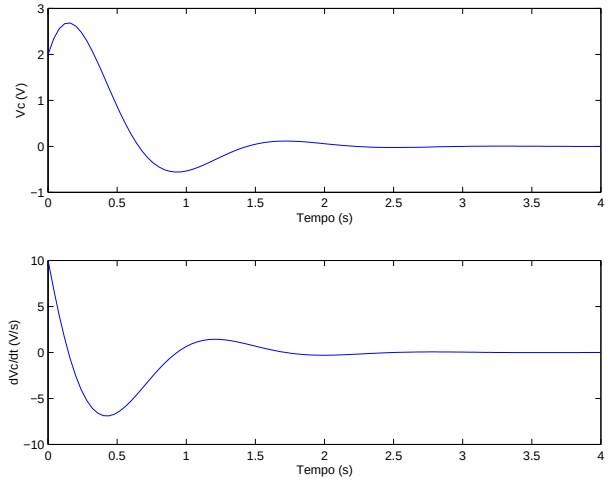


Figura 8.4: Resposta do circuito RLC a entrada $u(t) = 0$.

8.1.3 Supensão de Automóvel (Simplificada)

A suspensão de uma roda de veículo automator pode ser representada, de forma simplificada, pela massa M (Kg) suportada pela roda, um conjunto de molas representado pela mola ideal com constante K (N/m) e um amortecedor representado pelo sistema de absorção B (Ns/m). A Figura 8.5 apresenta os diversos componentes do sistema de suspensão. Conforme eixos coordenados, o sistema está em repouso na posição $y = 0$ e velocidade $\dot{y} = 0$.

O sistema de suspensão é submetido a uma força externa $f(t)$ dependente do terreno e da carga do veículo. As forças e respectivas direções de referência estão indicadas na figura. De acordo com a lei de Newton, a soma das forças que atuam no sistema deve igualar a massa vezes a aceleração, $\sum_{i=1}^{n} F_i = Ma$, ou seja, $f - F_M - F_K - F_B = Md^2y(t)/dt^2$. Formalmente:

$$f(t) - Mg - Ky(t) - B\frac{dy(t)}{dt} = M\frac{d^2y(t)}{dt^2} \tag{8.17}$$

$$\frac{d^2y(t)}{dt^2} + \frac{B}{M}\frac{dy(t)}{dt} + \frac{K}{M}y(t) = -g + \frac{1}{M}f(t) \tag{8.18}$$

Da mesma forma que no circuito RLC, vamos definir o estado do sistema como $x(t)$, sendo este dado por:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} y(t) \\ dy(t)/dt \end{bmatrix} \tag{8.19}$$

Procedendo à mudança de variável, substituímos $x(t)$ no lugar de $y(t)$ e $dy(t)/dt$ em (8.17)–(8.18), obtendo:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} y(t) \\ dy(t)/dt \end{bmatrix} \Leftrightarrow \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} \dot{y}(t) \\ \ddot{y}(t) \end{bmatrix} \tag{8.20}$$

que nos leva a:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} x_2(t) \\ -\frac{B}{M}\frac{dy(t)}{dt} - \frac{K}{M}y(t) - g + \frac{1}{M}f(t) \end{bmatrix} \\ &= \begin{bmatrix} x_2(t) \\ -\frac{B}{M}x_2(t) - \frac{K}{M}x_1(t) - g + \frac{1}{M}f(t) \end{bmatrix} \end{aligned} \tag{8.21}$$

Separando as influências do estado e externas, o sistema (8.21) assume a forma:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -K/M & -B/M \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/M \end{bmatrix} u + \begin{bmatrix} 0 \\ -1 \end{bmatrix} g \tag{8.22}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{8.23}$$

onde $u = f(t)$ é a entrada (força externa), $x(t)$ é o estado e $y = x_1(t)$ é a saída (posição).

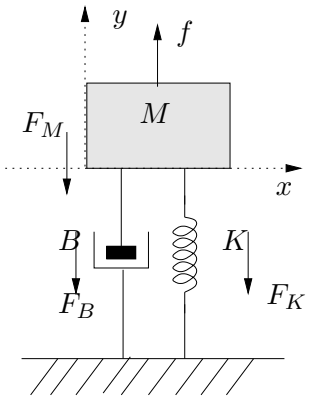


Figura 8.5: Suspensão de automóvel simplificada

8.1.4 Sistema de Massas Acopladas

O sistema de duas massas acopladas pode ser visto na Figura 8.6. Assumimos que a força exercida pela "*mola*" é nula quando os blocos estão separados de uma distância Δy, $F_K = 0$, enquanto a força exercida pelo "*amortecedor*" é nula se a variação de velocidade da massa M_1 em relação à M_2 é nula, $F_B = 0$. Aplicando a 2^a lei de Newton, a soma das forças aplicadas em cada massa iguala a massa vezes a aceleração, em notação matemática, isto equivale a dizer que:

$$\begin{aligned} M_1\dot{y}_1(t) &= F_K + F_B \\ &= K\left[y_2(t) - y_1(t) - \Delta_y\right] + B\left[\dot{y}_2(t) - \dot{y}_1(t)\right] \qquad (8.24) \\ M_2\dot{y}_2(t) &= f(t) - F_K - F_B \\ &= f(t) - K\left[y_2(t) - y_1(t) - \Delta_y\right] - B\left[\dot{y}_2(t) - \dot{y}_1(t)\right] \qquad (8.25) \end{aligned}$$

Deixando o vetor $x(t)$ definir as variáveis de estado como:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} = \begin{bmatrix} y_1(t) \\ \dot{y}_1(t) \\ y_2(t) \\ \dot{y}_2(t) \end{bmatrix}$$

podemos representar o sistema EDO de segunda ordem (8.24)–(8.25) como um sistema EDO de primeira ordem, $\dot{x} = Ax + Bu$, com variável de estado

dada pelo vetor x, como segue:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{K}{M_1} & -\frac{B}{M_1} & \frac{K}{M_1} & \frac{B}{M_1} \\ 0 & 0 & 0 & 1 \\ \frac{K}{M_2} & \frac{B}{M_2} & -\frac{K}{M_2} & -\frac{B}{M_2} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{K\Delta_y}{M_1} \\ 0 \\ \frac{K\Delta_y}{M_2} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{M_2} \end{bmatrix} u \tag{8.26}$$

com $u(t) = f(t)$.

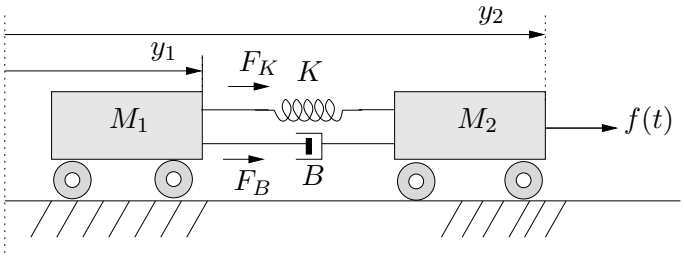


Figura 8.6: Sistema de duas massas acopladas

8.1.5 Motor de Corrente Contínua

Um modelo simplificado do motor de corrente contínua controlado pela armadura aparece na Figura 8.7. Neste esquema, $v_a(t)$ é a tensão aplicada à armadura, que está em série com o resistor R_a, o indutor L_a da armadura e a tensão $v_b(t)$ induzida pela corrente $i_a(t)$. A corrente da armadura gera um torque $T(t) = k_m i_a(t)$ proporcional à magnitude da corrente. O torque gerado movimenta a carga e o movimento rotacional produz a tensão $v_b(t)$ (*força eletromotriz*). O sistema ilustra a conversão de energia elétrica em energia mecânica. Sendo J o coeficiente de inércia da carga e D o coeficiente viscoso da mesma, temos pela 2^a lei de Newton que:

$$\begin{aligned} T(t) &= k_m i_a(t) \\ &= J\frac{d^2\Theta(t)}{dt^2} + D\frac{d\Theta(t)}{dt} \qquad (8.27) \\ v_b(t) &= k_b w(t) \\ &= k_b \frac{d\Theta(t)}{dt} \qquad (8.28) \end{aligned}$$

onde $w(t)$ é a velocidade angular. Equacionando o circuito da armadura, obtemos:

$$\begin{aligned} v_a(t) &= R_a i_a(t) + L_a \frac{di_a(t)}{dt} + v_b(t) \\ &= R_a i_a(t) + L_a \frac{di_a(t)}{dt} + k_b \frac{d\Theta(t)}{dt} \end{aligned} \tag{8.29}$$

Agora, escolhendo $x(t) = \left[\begin{array}{ccc} i_a(t) & \Theta(t) & d\Theta(t)/dt \end{array} \right]^T$ e $u(t) = v_a(t)$, podemos colocar as equações (8.27)–(8.29) na forma:

$$\left[\begin{array}{c} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{array} \right] = \left[\begin{array}{ccc} -R_a/L_a & 0 & -k_b/L_a \\ 0 & 0 & 1 \\ k_m/J & 0 & -D/J \end{array} \right] \left[\begin{array}{c} x_1(t) \\ x_2(t) \\ x_3(t) \end{array} \right] + \left[\begin{array}{c} 1/L_a \\ 0 \\ 0 \end{array} \right] u(t) \tag{8.30}$$

Note que (8.29) é um sistema de equações diferenciais ordinárias de primeira ordem.

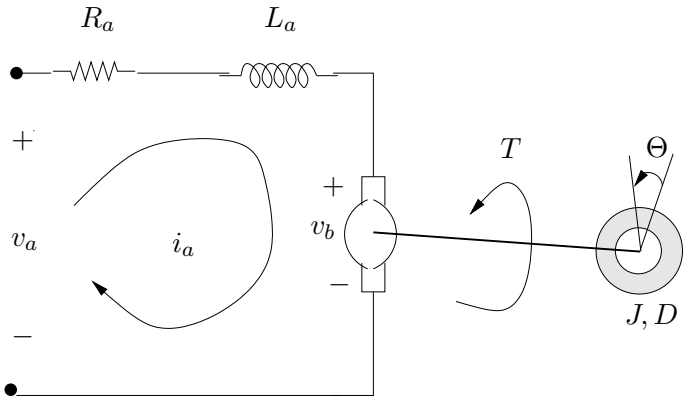


Figura 8.7: Motor de corrente contínua (CC) controlado pela armadura

8.1.6 Satélite em Órbita

Aqui consideramos a dinâmica de um satélite em órbita em torno da Terra, como mostra a Figura 8.8. Adota-se o sistema de coordenadas polares, onde $r(t)$ é a distância entre o satélite e a Terra e $\alpha(t)$ é o ângulo em relação à referência. Tanto o ângulo quanto a distância variam no tempo, portanto o movimento do satélite é caracterizado por $r(t)$, $\dot{r}(t)$, $\alpha(t)$ e $\dot{\alpha}(t)$. O satélite está equipado com um sistema de propulsão que produz uma impulsão na direção tangencial à sua trajetória, $F_t(t)$, e uma impulsão ortogonal à trajetória, $F_r(t)$. Conforme diagrama da figura, a velocidade tangencial $v_t(t)$

está relacionada à velocidade angular pela equação:

$$v_t(t) = r(t)\dot{\alpha}(t)$$

Já a velocidade radial é precisamente a taxa de variação radial:

$$v_r(t) = \dot{r}(t)$$

A energia cinética total do sistema, $k(t)$, instantânea é dada por:

$$\begin{aligned} k(t) &= \frac{1}{2}[v_t(t)^2 + v_r(t)^2] \\ &= \frac{1}{2}[r(t)^2\dot{\alpha}(t)^2 + \dot{r}(t)^2] \end{aligned} \tag{8.31}$$

Conforme o operador Lagrange, temos que:

$$\frac{d}{dt}\left[\frac{\partial k(t)}{\partial \dot{r}(t)}\right] - \left[\frac{\partial k(t)}{\partial r(t)}\right] = F_r(t) - \frac{cM}{r(t)^2} \tag{8.32}$$

$$\frac{d}{dt}\left[\frac{\partial k(t)}{\partial \dot{\alpha}(t)}\right] - \left[\frac{\partial k(t)}{\partial \alpha(t)}\right] = F_t(t) \tag{8.33}$$

onde c é a constante gravitacional. Desenvolvendo (8.32), concluímos que:

$$M\ddot{r}(t) - Mr(t)\dot{\alpha}(t)^2 = F_r(t) - \frac{cM}{r(t)^2}$$

ou, alternativamente,

$$\ddot{r}(t) = r(t)\dot{\alpha}(t)^2 - \frac{c}{r(t)^2} + F_r(t) \tag{8.34}$$

Similarmente, desenvolvendo (8.33), concluímos que:

$$2Mr(t)\dot{r}(t)\dot{\alpha}(t) + Mr(t)^2\ddot{\alpha}(t) = F_t(t)$$

ou

$$\ddot{\alpha}(t) = -\frac{2\dot{r}(t)\dot{\alpha}(t)}{r(t)} + \frac{F_t(t)}{Mr(t)^2} \tag{8.35}$$

Em síntese, a dinâmica do satélite em órbita é caracterizada pelas equações ordinárias de segunda ordem (8.34)–(8.35):

$$\ddot{r}(t) = r(t)\dot{\alpha}(t)^2 - \frac{c}{r(t)^2} + F_r(t) \tag{8.36}$$

$$\ddot{\alpha}(t) = -\frac{2\dot{r}(t)\dot{\alpha}(t)}{r(t)} + \frac{F_t(t)}{Mr(t)^2} \tag{8.37}$$

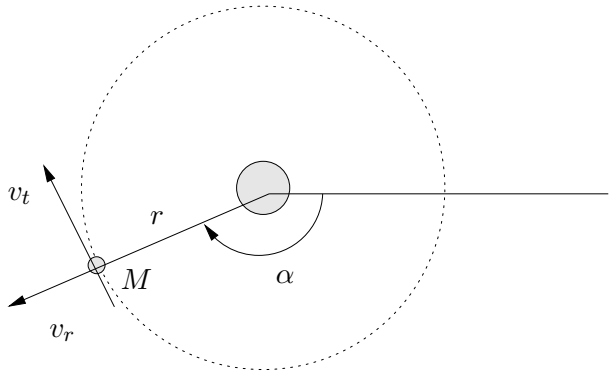


Figura 8.8: Satélite em órbita

8.1.7 Pêndulo Invertido

Aqui, ilustramos a concepção do sistema dinâmico que caracteriza o movimento de um veículo com pêndulo invertido acoplado, conforme mostra a Figura 8.9. Serão obtidas as equações diferenciais que regem o movimento do veículo e do pêndulo em resposta a forças externas e a ação da gravidade.

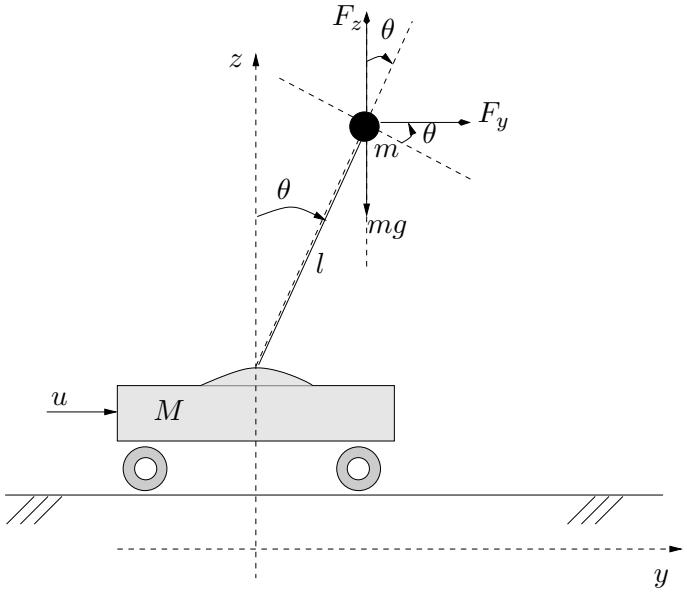


Figura 8.9: Veículo com pêndulo invertido

Sistema Dinâmico

Para tornar o desenvolvimento mais simples, assumimos que o carro e o pêndulo se movem no mesmo plano, e que podemos desprezar a fricção, a massa da haste e a força do vento. O problema clássico de controle consiste em encontrar uma lei de controle que mantenha o pêndulo na posição vertical, o qual se encontra deslocado da posição podendo estar se deslocando para baixo. Faz-se então uso da força horizontal para trazer o pêndulo de volta à posição vertical. A força horizontal é denotada por $u(t)$, a posição do horizontal do veículo é dada por $y(t)$ enquanto que a posição horizontal da massa do pêndulo é denotada por $y_p(t)$, e o ângulo da haste do pêndulo em relação ao eixo vertical é denotado por $\theta(t)$. Assumimos que no sistema de coordenadas (y, z) a origem de z é a posição onde a haste está acoplada ao carro, ou seja, $z = 0$ na origem do ângulo θ.

De acordo com as leis de Newton, as forças aplicadas segundo o eixo horizontal devem estar em equilíbrio, ou seja, a massa do carro multiplicada pela aceleração acrescida da massa do pêndulo multiplicada por sua aceleração deve igualar a força externa. Matematicamente, este princípio leva à equação:

$$M\frac{d^2}{dt^2}y + m\frac{d^2}{dt^2}y_p = u \tag{8.38}$$

A posição da massa do pêndulo pode ser expressa como uma função de y e do ângulo θ:

$$\begin{aligned} y_p &= y + l\sin\theta \\ z_p &= l\cos\theta \end{aligned} \tag{8.39}$$

onde l é o comprimento da haste do pêndulo. E substituindo (8.40) em (8.38) obtemos:

$$M\frac{d^2}{dt^2}y + m\frac{d^2}{dt^2}(y + l\sin\theta) = u \tag{8.40}$$

Observando que:

$$\frac{d}{dt}\sin\theta = (\cos\theta)\dot{\theta} \tag{8.41}$$

$$\frac{d^2}{dt^2}\sin\theta = -(\sin\theta)\dot{\theta}^2 + (\cos\theta)\ddot{\theta} \tag{8.42}$$

$$\frac{d}{dt}\cos\theta = -(\sin\theta)\dot{\theta} \tag{8.43}$$

$$\frac{d^2}{dt^2}\cos\theta = -(\cos\theta)\dot{\theta}^2 - (\sin\theta)\ddot{\theta} \tag{8.44}$$

podemos colocar (8.40) na forma:

$$(M + m)\ddot{y} - ml(\sin\theta)\dot{\theta}^2 + ml(\cos\theta)\ddot{\theta} = u \tag{8.45}$$

Aplicando agora as leis de Newton ao movimento rotacional, verificamos que os torques aplicados à massa do pêndulo devem estar em equilíbrio, ou seja, o torque resultante da aceleração angular deve igualar ao torque resultante da ação da gravidade. Isto leva à equação

$$(F_y \cos\theta)l - (F_z \sin\theta)l = (mg \sin\theta)l \tag{8.46}$$

As forças F_y e F_z são as componentes das forças que atuam na massa do pêndulo, podendo ser obtidas a partir das equações (8.39) e (8.41)–(8.44), resultando nas equações:

$$\begin{aligned} F_y &= m\frac{d^2}{dt^2}y_p \\ &= m\frac{d^2}{dt^2}(y + l\sin\theta) \\ &= m\left(\ddot{y} - l\sin\theta\dot{\theta}^2 + l\cos\theta\ddot{\theta}\right) \qquad (8.47) \\ F_z &= m\frac{d^2}{dt^2}z_p \\ &= -m\left(l\cos\theta\dot{\theta}^2 + l\sin\theta\ddot{\theta}\right) \qquad (8.48) \end{aligned}$$

Substituindo (8.47) e (8.48) em (8.46) e observando que l se cancela, obtemos a equação:

$$\begin{aligned} mg\sin\theta &= m\left(\ddot{y} - l\sin\theta\dot{\theta}^2 + l\cos\theta\ddot{\theta}\right)\cos\theta + m\left(l\cos\theta\dot{\theta}^2 + l\sin\theta\ddot{\theta}\right)\sin\theta \\ &= m\ddot{y}\cos\theta - ml\sin\theta\cos\theta\dot{\theta}^2 + ml\cos\theta\cos\theta\ddot{\theta} \\ &\quad + ml\cos\theta\sin\theta\dot{\theta}^2 + ml\sin\theta\sin\theta\ddot{\theta} \\ &= m\ddot{y}\cos\theta + ml\ddot{\theta} \qquad (8.49) \end{aligned}$$

Portanto, as equações que descrevem o sistema são (8.45) e (8.49), que justapostas levam ao sistema:

$$\begin{cases} (M+m)\ddot{y} - ml(\sin\theta)\dot{\theta}^2 + ml(\cos\theta)\ddot{\theta} = u \\ m\ddot{y}\cos\theta + ml\ddot{\theta} = mg\sin\theta \end{cases} \tag{8.50}$$

O sistema de equações diferenciais (8.50) é relativamente complexo em virtude de sua natureza não-linear.

Modelo Matricial

No que segue, modificamos o sistema (8.50) de maneira a deixar as variáveis $\ddot{y}$ e $\ddot{\theta}$ em função das demais. O sistema (8.50) é equivalente a:

$$\begin{aligned} (M+m)\ddot{y} + ml(\cos\theta)\ddot{\theta} &= u + ml(\sin\theta)\dot{\theta}^2 \\ m\ddot{y}\cos\theta + ml\ddot{\theta} &= mg\sin\theta \end{aligned}$$

que pode ser colocado em forma matricial:

$$\begin{bmatrix} (M+m) & ml\cos\theta \\ m\cos\theta & ml \end{bmatrix} \begin{bmatrix} \ddot{y} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} u + ml\sin\theta\dot{\theta}^2 \\ mg\sin\theta \end{bmatrix} \tag{8.51}$$

E multiplicando pela matriz inversa ambos os lados da igualdade (8.51) obtemos:

$$\begin{aligned} \begin{bmatrix} \ddot{y} \\ \ddot{\theta} \end{bmatrix} &= \frac{1}{(M+m)ml - m^2 l\cos^2\theta} \begin{bmatrix} ml & -ml\cos\theta \\ -m\cos\theta & (M+m) \end{bmatrix} \begin{bmatrix} u + ml\sin\theta\dot{\theta}^2 \\ mg\sin\theta \end{bmatrix} \\ &= \frac{1}{(M+m)l - ml\cos^2\theta} \begin{bmatrix} lu + ml^2\sin\theta\dot{\theta}^2 - mlg\sin\theta\cos\theta \\ -\cos\theta u - ml\sin\theta\cos\theta\dot{\theta}^2 + (M+m)g\sin\theta \end{bmatrix} \end{aligned} \tag{8.52}$$

Sistema de Equações de 1^a Ordem

Podemos expressar (8.52) em um sistema envolvendo apenas derivadas de variáveis por meio de uma mudança de variáveis, fazendo:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y \\ \dot{y} \\ \theta \\ \dot{\theta} \end{bmatrix} \Rightarrow \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} \dot{y} \\ \ddot{y} \\ \dot{\theta} \\ \ddot{\theta} \end{bmatrix} \tag{8.53}$$

o que nos leva a concluir que:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{u + ml\sin x_3 {x_4}^2 - mg\sin x_3\cos x_3}{(M+m) - m\cos^2 x_3} \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= \frac{-\cos x_3 u - ml\sin x_3\cos x_3 {x_4}^2 + (M+m)g\sin x_3}{(M+m)l - ml\cos^2 x_3} \end{aligned} \tag{8.54}$$

Concluímos que o sistema pode ser modelado pelas equações (8.50) ou, equivalentemente, pelas equações (8.54) por meio da mudança de variáveis definida por (8.53). De forma mais compacta, (8.54) pode ser colocado em forma vetorial fazendo $x = (x_1, x_2, x_3, x_4)$ e $F(x, u)$ corresponder ao lado direito do sistema (8.54), o que leva à:

$$\dot{x} = F(x, u) \tag{8.55}$$

Linearização

Podemos, alternativamente, considerar uma linearização do sistema (8.50) em torno do ponto de equilíbrio $\theta = 0$ e $\dot{\theta} = 0$, colocando o pêndulo na posição vertical, o que será objeto de análise mais à frente.

Tomando como ponto de equilíbrio a posição $(y, \dot{y}, \theta, \dot{\theta}) = 0$ e força $u = 0$ e assumindo pequenas perturbações no ponto de equilíbrio[1], podemos simplificar o modelo $\dot{x} = F(x, u)$ por meio de uma aproximação linear fazendo uso da série de Taylor:

$$\begin{aligned} \dot{x} &= F(0,0) + \nabla_x F(0,0)x + \nabla_u F(0,0)u \\ &= \nabla_x F(0,0)x + \nabla_u F(0,0)u \end{aligned} \tag{8.56}$$

Fazendo $F(x,u) = [f_1, f_2, f_3, f_4]^T$, com $f_1 = x_2$, $f_2 = (u + ml\sin x_3 {x_4}^2 - mg\sin x_3\cos x_3)/[(M+m) - m\cos^2 x_3]$, $f_3 = x_4$ e $f_4 = (-\cos x_3 u - ml\sin x_3\cos x_3 {x_4}^2 + (M+m)g\sin x_3)/[(M+m)l - ml\cos^2 x_3]$, podemos proceder a linearização, conforme segue. Para f_1, obtemos:

$$\frac{\partial}{\partial x_1} f_1 = 0,\ \frac{\partial}{\partial x_2} f_1 = 1,\ \frac{\partial}{\partial x_3} f_1 = 0,\ \frac{\partial}{\partial x_4} f_1 = 0,\ \frac{\partial}{\partial u} f_1 = 0$$

Para f_2, obtemos:

$$\frac{\partial}{\partial x_1} f_2 = 0,\ \frac{\partial}{\partial x_2} f_2 = 0$$

$$\begin{aligned} \frac{\partial}{\partial x_3} f_2 &= \frac{ml\cos x_3 {x_4}^2 - mg\cos^2 x_3 + mg\sin^2 x_3}{(M+m) - m\cos^2 x_3} \\ &\quad - \frac{(u + ml\sin x_3 {x_4}^2 - mg\sin x_3\cos x_3)(2m\cos x_3 \sin x_3)}{[(M+m) - m\cos^2 x_3]^2} \\ \frac{\partial}{\partial x_3} f_2(0) &= -\frac{mg}{M} \\ \frac{\partial}{\partial x_4} f_2 &= \frac{2ml\sin x_3 x_4}{(M+m) - m\cos^2 x_3} \ \Rightarrow\ \frac{\partial}{\partial x_4} f_2(0) = 0 \\ \frac{\partial}{\partial u} f_2 &= \frac{1}{(M+m) - m\cos^2 x_3} \ \Rightarrow\ \frac{\partial}{\partial u} f_2(0) = \frac{1}{M} \end{aligned}$$

Para f_3, obtemos:

$$\frac{\partial}{\partial x_1} f_3 = 0,\ \frac{\partial}{\partial x_2} f_3 = 0,\ \frac{\partial}{\partial x_3} f_3 = 0,\ \frac{\partial}{\partial x_4} f_3 = 1,\ \frac{\partial}{\partial u} f_3 = 0,$$

[1] Note que $F(0,0) = 0$ o que implica $\dot{x} = 0$

Para f_4, obtemos:

$$\frac{\partial}{\partial x_1} f_4 = 0, \ \frac{\partial}{\partial x_2} f_4 = 0$$

$$\begin{aligned}
\frac{\partial}{\partial x_3} f_4 &= \frac{\sin x_3 u - ml\cos^2 x_3 x_4^2 + ml\sin^2 x_3 x_4^2 + (M+m)g\cos x_3}{(M+m)l - ml\cos^2 x_3} \\
&\quad - \frac{[2ml\cos x_3 \sin x_3][-\cos x_3 u - ml\sin x_3 \cos x_3 {x_4}^2 + (M+m)g\sin x_3]}{[(M+m)l - ml\cos^2 x_3]^2} \\
\frac{\partial}{\partial x_3} f_4(0) &= \frac{(M+m)g}{Ml} \\
\frac{\partial}{\partial x_4} f_4 &= \frac{-2ml\sin x_3 \cos x_3 x_4}{(M+m)l - ml\cos^2 x_3} \ \Rightarrow \ \frac{\partial}{\partial x_4} f_4(0) = 0 \\
\frac{\partial}{\partial u} f_4 &= \frac{-\cos x_3}{(M+m)l - ml\cos^2 x_3} \ \Rightarrow \ \frac{\partial}{\partial u} f_4(0) = -\frac{1}{Ml}
\end{aligned}$$

A partir das derivadas $\partial f_i(0)/\partial x_j$, geramos a aproximação linear:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{(M+m)g}{Ml} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ -\frac{1}{Ml} \end{bmatrix} u \qquad (8.57)$$

Problemas como o descrito acima tem as relações entre as variáveis descritas em termos de equações diferenciais, ou seja, equações que envolvem um função desconhecida e algumas de suas derivadas. Uma equação que envolve derivadas até ordem n é chamada de equação diferencial ordinária (ODE).

8.2 Exemplos de Equações Diferenciais

Exemplo

Uma lista de equações diferenciais exemplo segue abaixo:

a) $\frac{dy}{dx} = y + x^2$

b) $\frac{dy}{dx} = y^2$

c) $\frac{dy}{dx} = 2x + 3$

d) $e^x \frac{dy}{dx} + 7xy = x^2 + 1$

e) $\frac{dy}{dx} = y + 1$

f) $\frac{d^2y}{dx^2} + 3\frac{dy}{dx} - 17y = 0$

g) $xyy'' + xy' = 0$

h) $e^x y'' + 2y' + 3xy = x + 3$

As equações dadas em (a) e (e) são equações diferenciais de primeira ordem e lineares. Já as equações (f) e (h) são equações diferenciais de segunda ordem e lineares, enquanto que a equação (g) é uma equação diferencial de segunda ordem e não-linear.

Exemplo

Considere a equação diferencial linear de primeira ordem:

$$\frac{dy}{dx} = 2x + 3$$

que pode ser escrita como $y' = f(x)$ sendo $f(x) = 2x+3$ uma função contínua para $a < x < b$. A solução da equação é dada por:

$$\begin{aligned} y &= \int f(x)dx + c \\ &= \int (2x+3)dx \\ &= x^2 + 3x + c \end{aligned}$$

8.3 Problema de Valor Inicial

O problema de valor inicial consiste em encontrar uma solução para a equação diferencial

$$y^{(n)}(x) = f(x, y, y', \ldots, y^{(n-1)}) \tag{8.58}$$

sendo as condições iniciais dadas por:

$$\begin{aligned} y(a) &= \xi_1 \\ y'(a) &= \xi_2 \\ \vdots &\quad \vdots \\ y^{(n-1)}(a) &= \xi_n \end{aligned}$$

No que segue desenvolveremos métodos numéricos para resolver of problema (8.58).

8.4 Sistemas de Equações Diferenciais

Um sistema de equações diferenciais de primeira ordem tem a seguinte forma:

$$\begin{cases} y_1'(x) &= f_1(x, y_1, y_2, \ldots, y_n) \\ y_2'(x) &= f_2(x, y_1, y_2, \ldots, y_n) \\ \quad\vdots & \\ y_n'(x) &= f_n(x, y_1, y_2, \ldots, y_n) \end{cases} \tag{8.59}$$

Quando o problema acima tem solução, então ele tem, em geral, várias soluções, ou seja, uma família de soluções. Com as condições iniciais abaixo, temos o problema do valor inicial:

$$\begin{cases} y_1(x) &= y(x) \\ y_2(x) &= y'(x) \\ \quad\vdots & \\ y_n(x) &= y^{n-1}(x) \end{cases}$$

Note que a equação diferencial do problema de valor inicial, (8.58), pode ser colocada na forma de um sistema de equações diferenciais de primeira ordem, conforme modelo dado por (8.59). Para tanto, basta proceder como segue:

$$\begin{aligned} y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \quad &\Leftrightarrow \quad \begin{cases} y^{(n)}(x) &= f(x, y, y', \ldots, y^{(n-1)}) \\ y_1 &= y \\ y_2 &= y' \\ \quad\vdots & \\ y_n &= y^{(n-1)} \end{cases} \\ &\Leftrightarrow \quad \begin{cases} y_1' &= y' \\ y_2' &= y'' \\ \quad\vdots & \\ y_n' &= y^{(n)} \end{cases} \\ &\Leftrightarrow \quad \begin{cases} y_1' &= y_2 \\ y_2' &= y_3' \\ \quad\vdots & \\ y_n' &= f(x, y_1, y_2, \ldots, y_n) \end{cases} \end{aligned} \tag{8.60}$$

Exemplo

Considere o problema de valor inicial dado por:

$$y'''(x) = xy' + e^x y(x) + x^2 + 1, \ 0 \leqslant x < 1 \tag{8.61}$$

tal que $y(0) = 1$, $y'(0) = 0$, e $y''(0) = -1$. Podemos então transformar (8.61) em um sistema de equações de primeira ordem, fazendo:

$$\left\{\begin{array}{lcl} y_1 &=& y \\ y_2 &=& y' \\ y_3 &=& y'' \end{array}\right. \Rightarrow \left\{\begin{array}{l} y_1' = y_2 \\ y_2' = y_3 \\ y_3' = xy_2 + e^x y_1 + x^2 + 1 \\ y_1(0) = 1,\ y_2(0) = 0,\ y_3(0) = -1 \end{array}\right. \tag{8.62}$$

8.5 Equações de Diferenças

Uma equação de diferenças de ordem n é uma sequência de equações da forma:

$$\begin{array}{rcl} g_k(y_{k+n}, y_{k+n-1}, \ldots, y_k) &=& 0,\ k = 0, 1, 2 \ldots \\ y_j &=& \xi_j,\ j = 0, 1, 2 \ldots, n-1 \end{array} \tag{8.63}$$

Os g_k são funções de $n+1$ variáveis e os valores ξ_j são dados específicos. Uma solução de tal equação é uma sequência $(y_0, y_1, \ldots, y_{n-1}, y_n, y_{n+1}, \ldots)$ que satisfaz as equações (8.63).

Uma forma especial das equações (8.63) é:

$$\begin{array}{rcl} \alpha_n y_{k+n} + \alpha_{n-1} y_{k+n-1} + \ldots + \alpha_{y_k} &=& 0,\ k = 0, 1, 2, \ldots \\ y_j &=& \xi_j,\ j = 0, 1, 2, \ldots, n-1 \end{array} \tag{8.64}$$

Em (8.64), os g_k independem de k e são funções lineares homogêneas de todas as variáveis, e por esta razão são chamadas de equações de diferenças lineares homogêneas, com coeficientes constantes.

Exemplos

Abaixo listamos três exemplos de equações de diferenças lineares:

a) $y_{k+2} - 5y_{k+1} + 6y_k = 0$, $y_0 = 0$, $y_1 = 1$

b) $y_{k+1} - y_k = 0$ e $y_0 = 0$

c) $y_{k+3} - 2y_{k+2} - y_{k+1} + 2y_k = 0$, $y_0 = 0$, $y_1 = -3$, e $y_2 = 1$

8.6 Método de Euler

Estudaremos agora métodos que aproximam uma equação diferencial por uma equação de diferenças. Determinar numericamente uma solução de uma

equação diferencial é encontrar os valores $(y_1, y_2, \ldots, y_n)$ através de uma aproximação da equação de diferenças. Tal aproximação introduz um erro de truncamento e um erro de arredondamento.

Vamos resolver a ODE de primeira ordem da forma $y' = f(x, y)$ sujeita à condição inicial $y(x_0) = y_0$. Primeiramente, vamos analisar o problema graficamente. Suponhamos que $y = F(x)$ e que a solução analítica seja a curva mostrada no gráfico da figura abaixo.

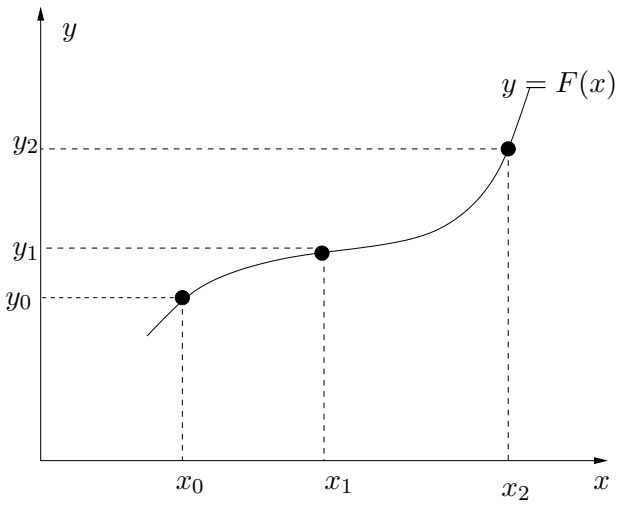


Figura 8.10: Ilustração de uma primitiva $F(x)$

Para fazer uma estimativa de y_1, em torno do ponto (x_0, y_0) vamos considerar que:

$$\frac{dy}{dx}|_{(x_0,y_0)} = f(x_0, y_0)$$

que pode ser aproximado em torno de (x_0, y_0) por:

$$\frac{y - y_0}{x - x_0} \cong f(x_0, y_0)$$

Observando que, se $h = x_1 - x_0$ tender a zero, a ordenada do ponto Q $(\bar{y})$ tende a y_1 e daí:

$$\begin{array}{rcl} \bar{y} & = & y_0 + hf(x_0, y_0) \\ y_1 & \cong & y_0 + hf(x_0, y_0) \end{array} \tag{8.65}$$

Generalizando, temos a seguinte equação a diferenças, que é a expressão de Euler:

$$y_{k+1} = y_k + hf(x_k, y_k) \tag{8.66}$$

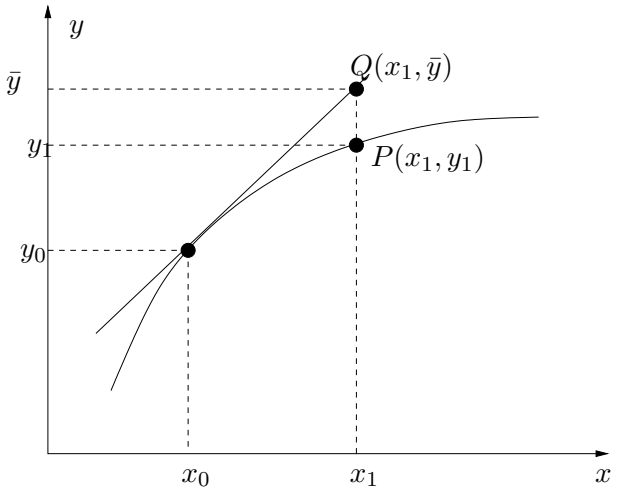


Figura 8.11: Método de Euler

Outro enfoque consiste em considerar a aproximação:

$$y'(x) \cong \frac{[y(x+h) - y(x)]}{h} \tag{8.67}$$

Introduzindo a notação

$$x_k = a + kh, \quad = 0, 1, 2, \ldots$$

de modo que $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. Fazendo y_k representar uma aproximação para $y(x_k)$ onde $y(x)$ é a solução de $y' = f(x, y(x))$, então (8.67) sugere que:

$$y'(x_k) = \frac{[y_{k+1} - y_k]}{h}$$

Portanto,

$$y_{k+1} - y_k = hy'(x_k) \Rightarrow y_{k+1} = y_k + hf(x_k, y_k) \tag{8.68}$$

que é novamente a expressão do método de Euler.

8.6.1 Exemplo

Resolver a equação diferencial $y' = 2x + 3$, para $x = \{1.0, 1.1, 1.2, 1.3\}$, tendo como condições iniciais $y = 1$ quando $x = 1$.

Temos que $f(x, y) = 2x + 3$, $x_0 = 1$, $y_0 = 1$, $h = 0.1$.

Passo 0 Temos, pelas condições iniciais que:

$$\begin{aligned} y_0 &= 1 \\ x_0 &= 1 \end{aligned}$$

Passo 1 Calculamos y_1 para $x_1 = 1.1$ conforme segue:

$$\begin{aligned} y_1 &= y_0 + hf(x_0, y_0) \\ &= 1 + 0.1(2 \times 1 + 3) \\ &= 1.5 \\ x_1 &= x_0 + h \\ x_0 &= 1.0 y_0 = 1.0 \\ x_1 &= 1.1 y_1 = 1.5 \end{aligned}$$

Passo 2 Calculamos y_2 para $x_2 = 1.2$ conforme segue:

$$\begin{aligned} y_2 &= y_1 + hf(x_1, y_1) \\ &= 1.5 + 0.1(2 \times 1.1 + 3) \\ &= 2.02 \\ x_2 &= 1.2 \end{aligned}$$

Passo 3 Calculamos y_3 para $x_3 = 1.3$ conforme segue:

$$\begin{aligned} y_3 &= y_2 + hf(x_2, y_2) \\ &= 2.56 \\ x_3 &= 1.3 \end{aligned}$$

8.6.2 O Algoritmo de Euler

Algoritmo de Euler (f, a, b, ξ, h)

$x \leftarrow a$
$y \leftarrow \xi$
`Enquanto` $x \leqslant b$
    $y \leftarrow y + hf(x, y)$
    $x \leftarrow x + h$
    `Saida(x,y)`
`Fim enquanto`

8.7 Método de Euler para Sistemas de Equações

Aqui vamos estender o método de Euler desenvolvido na seção para resolver numericamente sistemas de equações diferenciais.. Vejamos inicialmente

um sistema de duas equações:

$$\begin{cases} y' = f(t, y, z) \\ z' = g(t, y, z) \end{cases}$$

Fazendo $h_0 = t_1 - t_0$, temos:

$$\begin{cases} y(t_1) &= y_1 \\ &= y_0 + h_0 f(t_0, y_0, z_0) \\ z(t_1) &= z_1 \\ &= z_0 + h_0 g(t_0, y_0, z_0) \end{cases}$$

Generalizando para um passo h qualquer, temos:

$$\begin{cases} y_{k+1} &= y_k + hf(t_k, y_k, z_k) \\ z_{k+1} &= z_k + hg(t_k, y_k, z_k) \end{cases}$$

Para um conjunto de n equações o método assume a forma:

$$\begin{cases} y_1^{k+1} &= y_1^k + hf_1(x(k), y_1^k, y_2^k, \ldots, y_n^k) \\ y_2^{k+1} &= y_2^k + hf_2(x(k), y_1^k, y_2^k, \ldots, y_n^k) \\ \quad\vdots & \\ y_n^{k+1} &= y_n^k + hf_n(x(k), y_1^k, y_2^k, \ldots, y_n^k) \end{cases} \tag{8.69}$$

Logo, as equações (8.69) nos dão um processo iterativo para calcular a solução numérica aproximada de $y(x)$ a partir de um conjunto de condições iniciais. Isto significa que a trajetória $\{(x_k, y^k) : k = 0, 1, 2, \ldots\}$, conforme (8.69), produz uma solução aproximada para $y(x)$.

8.8 Métodos Baseados na Série de Taylor

Tomemos como ponte de partida a equação diferencial ordinária:

$$\begin{aligned} y' &= f(x, y(x)) \\ y(x_0) &= y_0 \end{aligned}$$

Seja $y = F(x)$ a solução, ou seja, $F'(x) = f(x, y)$ com $F(x_0) = y_0$. Assumiremos que F é diferenciável até ordem n. Expandindo $F(x)$ na série de Taylor em torno de x_0, temos:

$$\begin{aligned} F(x) &= F(x_0) + \frac{x - x_0}{1!} F'(x_0) + \frac{(x - x_0)^2}{2!} F''(x_0) + \ldots \\ &\quad + \frac{(x - x_0)^n}{n!} F^{(n)}(x_0) \end{aligned} \tag{8.70}$$

Sabemos que:

$$\begin{aligned} F(x_0) &= y_0 \\ F'(x_0) &= f(x_0, y_0) \\ &= y'_0 \end{aligned}$$

Entretanto precisamos ainda determinar y''_0, $y'''_0, \ldots, y_0^{(n)}$. Não conhecemos estas derivadas pois $F(x)$ não é conhecida. Se f for suficientemente derivável, elas podem ser determinadas considerando-se a derivada total em relação a x, pois f é função implícita de y. Isto nos leva aos desenvolvimento abaixo:

$$\begin{aligned} y' &= f(x, y(x)) \\ \\ y'' &= f'(x, y(x)) \\ &= \tfrac{\partial f}{\partial y}\tfrac{dy}{dx} + \tfrac{\partial f}{\partial x} \\ &= f_y(x, y(x)).f(x, y(x)) + f_x(x, y(x)) \\ &= f_y f + f_x \\ \\ y''' &= \tfrac{\partial f_y f}{\partial y}.\tfrac{dy}{dx} + \tfrac{\partial f_y f}{\partial x} + \tfrac{\partial f_x}{\partial x}.\tfrac{dy}{dx}.\tfrac{\partial f_x}{\partial x} \\ &= f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_x f_y + f_y^2 f \end{aligned} \tag{8.71}$$

8.8.1 Exemplo

Considere a ODE $y' = x + y^2$ e a condição inicial $y(0) = 1$. Conforme desenvolvimento acima, sabemos que

$$y'' = f_x + f_y f = 1 + 2y'y.$$

Observando que $f_x = 1$, $f_y = 2y$, $f_{xx} = 0$ e $f_{xy} = 0$, podemos verificar que:

$$\begin{aligned} y''' &= f_{yy}f^2 + f_x f_y + f_y^2 f \\ &= 2(x + y^2)^2 + 1.(2y) + (2y)^2(x + y^2) \\ &= 2(x + 2xy^2 + y^4) + 2y + 4y^2 x + 4y^4 \\ &= 2x + 4xy^2 + 2y^4 + 2y + 4y^2 x + 4x^4 \\ &= 6y^4 + 2y + 4xy^2 + 4y^2 x + 2x \end{aligned}$$

Asim temos:

$$\begin{aligned} y'(0) &= 0 + y_0^2 = y_0^2 \\ &= 1 \end{aligned}$$

$$\begin{aligned} y''(0) &= 1 + 2yy' \\ &= 1 + 2y_0y_0' \\ &= 1 + 2.1.1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} y'''(0) &= 6y^4 + 2y + 4xy^2 + 4y^2x + 2x \\ &= 6y_0^4 + 2y_0 \\ &= 6.1 + 2.1 \\ &= 8 \end{aligned}$$

Isto nos leva à solução aproximada:

$$y(x) = 1 + x + \frac{3}{2}x^2 + \frac{8}{6}x^3 + E_T$$

Onde E_T denota o erro cometido.

8.9 Método de Runge-Kutta

Os métodos de Runge-Kutta são obtidos pela série de Taylor em que se omite os termos de mais alta ordem na expansão. Se cancelarmos os termos que contêm potências de h de ordem maior que p, obtemos um método de ordem p. O método de Euler estudado anteriormente é de primeira ordem. Para desenvolvermos os métodos, vamos expandir y_{k+1} em vez de $F(x)$ como descrito na equação (8.70), ou seja:

$$y_{k+1} = y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(x_k) + \frac{h^3}{6}y'''(x_k) \qquad (8.72)$$

8.9.1 Método de Runge-Kutta de Segunda Ordem

Assumindo que $y(x)$ é três vezes continuamente diferenciável, então o teorema de Taylor nos dá:

$$y_{k+1} = y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(x_k) + \frac{h^3}{6}y'''(\xi_k) \qquad (8.73)$$

para algum $\xi_k \in [x_k, x_{k+1}]$. Usando a notação

$$y'(x_k) = f(x_k, y(x_k))$$

vemos que

$$y(x_{k+1}) = y(x_k) + hf(x_k, y(x_k)) + \frac{h^2}{2}[\frac{df(x, y(x))}{dx}] + O(h^3) \quad (8.74)$$

Para calcular $\frac{df(x,y(x))}{dx}$ poderíamos usar uma das fórmulas de (8.71), mas teríamos o problema das derivadas parciais, então usamos uma aproximação dada pela derivação do polinômio interpolador de grau um, ou seja:

$$p(x) = f(x_1)\frac{x - x_2}{x_1 - x_2} + f(x_2)\frac{x - x_1}{x_2 - x_1}$$

então:

$$p'(x) = f(x_1)\frac{1}{x_1 - x_2} + f(x_2)\frac{1}{x_2 - x_1}$$

daí temos com $h = x_2 - x_1$ que:

$$\frac{df(x, y(x))}{dx} = \frac{1}{h}\left[f(x + h, y(x + h)) - f(x, y(x))\right] + O(h) \quad (8.75)$$

Escrevendo a equação (8.75) para $x = x_k$ e substituindo em (8.74), temos:

$$\begin{aligned} y(x_{k+1}) &= y(x_k) + hf(x_k, y(x_k)) \\ &\quad + \frac{h^2}{2}\left[\frac{f(x_{k+1}, y(x_{k+1})) - f(x_k, y(x_k))}{h}\right] + O(h^3) \\ &\cong y(x_k) + \frac{h}{2}\left[f(x_{k+1}, y(x_{k+1})) + f(x_k, y(x_k))\right] \\ &= y_k + \frac{h}{2}\left[f(x_{k+1}, y_{k+1}) - f(x_k, y_k)\right] \end{aligned} \quad (8.76)$$

Todavia, a fórmula (8.76) não pode ser utilizada para calcular $y_1, y_2, \ldots$ por causa do termo $y_{k+1} = y(x_{k+1})$ no lado direito da igualdade. As fórmulas do tipo (8.76) são chamadas fórmulas implícitas. Substituindo y_{k+1} em $f(x_{k+1}, y_{k+1})$ pela expressão do método de Euler, temos:

$$\begin{aligned} f(x_{k+1}, y_{k+1}) &= f(x_{k+1}, y(x_k) + hy'(x_k) + O(h^2)) \\ &= f(x_{k+1}, y(x_k) + hy'(x_k)) + O(h^2) \end{aligned} \quad (8.77)$$

A substituição de (8.77) em (8.76), nos leva a:

$$y_{k+1} = y_k + \frac{h}{2}[f(x_k, y_k) + f(x_{k+1}, y_k + hf(x_k, y_k)] + O(h^3) \quad (8.78)$$

que é conhecida como fórmula de Runge-Kutta de segunda ordem.

Para uma equação diferencial do tipo $y' = f(x, y)$, com condição inicial $y(x_0) = y_0$, O método Runge-Kutta de segunda-ordem pode ser reescrito da seguinte forma:

$$\begin{aligned} x_{k+1} &= x_k + h \\ y_{k+1} &= y_k + \frac{k_1 + k_2}{2}, \end{aligned}$$

onde:

$$\begin{aligned} k_1 &= hf(x_k, y_k) \\ k_2 &= hf(x_{k+1}, y_k + k_1) \end{aligned}$$

8.10 Exercícios

Exercício 8.1 Considere o Problema com Condições de Contorno a seguir:

$$-\frac{d^2x(t)}{dt^2} + x(t) = r(t) \quad , \quad \text{com} \begin{cases} t_0 \leqslant t \leqslant t_f \\ x(t_0) = x_0 \ \text{ e } x(t_f) = x_f \end{cases} \tag{8.79}$$

O problema, contínuo na variável t, consiste em determinar a função $x(t)$ que satisfaz (8.79).

Este problema pode ser tratado computacionalmente através da discretização do intervalo de tempo considerado. Como $r(t)$ é uma função conhecida, divide-se o intervalo $[t_0\,,\, t_f]$ em $n+1$ subintervalos igualmente espaçados e considera-se a informação relativa a $r(t)$ em n instantes de tempo discretos:

$$t_k = t_0 + kT \quad \Rightarrow \quad r(t_k) = r_k \ , \quad k = 1, 2, \ldots, n$$

Deseja-se, então, calcular os valores aproximados de $x(t)$ nos instantes determinados, levando em consideração as condições de contorno:

$$\begin{aligned} &x(t_k) = x_k \ , \quad k = 1, 2, \ldots, n \\ &\textit{com } x(t_0) = x_0 \ \textit{e } x(tf) = x\left(t_0 + (n+1)T\right) = x_f \end{aligned}$$

Para tanto, pode-se utilizar a aproximação seguinte:

$$\frac{d^2x(t)}{dt^2} \approx \frac{x(t+T) - 2x(t) + x(t-T)}{T^2} \tag{8.80}$$

a) Equacione a estratégia de discretização delineada acima, conhecida como *Método das diferenças finitas*, sob a forma de um sistema de equações lineares que permita determinar as aproximações x_k.

b) Considere $r(t) = t$, $t_0 = 0$, $t_f = 1$, $x(0) = x(1) = 0$ e $n = 6$:

i. Utilize um método direto para obter a solução do sistema de equações correspondente. Justifique a escolha de método.

ii. Utilize o Método de Gauss-Siedel para encontrar uma aproximação para a solução (10 iterações / aproximação inicial "nula"). Analise a convergência do método.

Exercício 8.2 Para o circuito dado na Figura 8.12, a equação diferencial ordinária de segunda ordem que descreve o comportamento dinâmico do sistema é dada por:

$$L\frac{d^2}{dt^2}i(t) + R\frac{d}{dt}i(t) + \frac{1}{C}i(t) = 3.5\cos(3.5t)$$

Executar as seguintes tarefas:

a) Calcular utilizando o método de Euler o valor da corrente após 3.8 segundos, assumindo que $i(0) = i'(0) = 0$. Utilizar passo de integração suficientemente pequeno. A implementação deve ser feita em Matlab, Octave ou Scilab.

b) Calcular utilizando o método de Runge-Kutta o valor da corrente após 3.8 segundos, assumindo que $i(0) = i'(0) = 0$. Utilizar passo de integração suficientemente pequeno. A implementação deve ser feita em Matlab, Octave ou Scilab.

c) Gerar gráfico da curva $i(t)$ para o período de integração.

d) Gerar gráfico da curva $i'(t)$ para o período de integração.

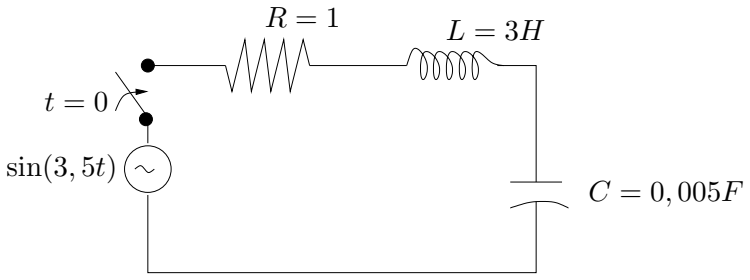


Figura 8.12: Circuito elétrico

Exercício 8.3 Para o pêndulo invertido ilustrado na Figura 8.9, a dinâmica do movimento pode ser aproximada em torno do estado $(y, \dot{y}, \theta, \dot{\theta}) = (0, 0, 0, 0)$ pelas equações diferenciais abaixo:

$$\begin{array}{rcl} (M+m)\ddot{y} + ml\ddot{\theta} & = & u \\ 2l\ddot{\theta} - 2g\theta + \ddot{y} & = & 0 \end{array} \tag{8.81}$$

onde $M = 10Kg$, $m = 0.5Kg$, $g = 9.8m/s^2$, e $l = 50cm$. Executar as seguintes tarefas:

a) Fazendo $x_1 = y, x_2 = \dot{y}, x_3 = \theta, x_4 = \dot{\theta}$ e $x = (x_1, x_2, x_3, x_4)$, obter um sistema de equações diferenciais

$$\dot{x} = F(x, u)$$

equivalente ao sistema de equações (8.81).

b) Seja a lei de controle dada por $u(t) = -Kx(t)$ onde

$$K = [\ -3.1623 \quad -10.1554 \quad -494.4166 \quad -110.6079\ ]$$

Obter a trajetória de $x(t)$ para $t \in [0, 10s]$ usando o método de Runge-Kutta e fazendo $x(0) = [\ 0.2 \quad 0 \quad 0.5 \quad 0\ ]^T$. Apresentar em gráficos as curvas $x_1(t), \ldots, x_4(t)$. Qual é o valor de $x(10)$?

c) Repetir o item (b) com $x(0) = [\ -0.3 \quad 0 \quad 0.8 \quad 0\ ]^T$.

Exercício 8.4 Considere o sistema mecânico da Figura 8.13, onde:

- $M = 5kg$, massa do corpo;
- $K = 3m/N$, coeficiente da mola;
- $D = 1Ns/m$, coeficiente de atrito viscoso;
- $F = 10e^{-2t}N$, força aplicada no corpo; e
- $V(t)$, velocidade do corpo.

A equação diferencial que caracteriza o movimento do corpo é dada por:

$$5\frac{d^2V(t)}{dt^2} + 4\frac{dV(t)}{dt} + \frac{1}{3}V(t) = -20e^{-2t}$$

Sabemos que para $t = 0s$ a velocidade do corpo é $1.5m/s$ e a aceleração é $0.5m/s^2$.

a) Calcular utilizando o método de Euler a velocidade e aceleração do corpo após $5s$. Utilizar passo de integração suficientemente pequeno. A implementação deve ser feita em Matlab, Octave ou Scilab.

b) Calcular utilizando o método de Runge-Kutta de 2a ordem a velocidade e aceleração do corpo após $5s$. Utilizar passo de integração suficientemente pequeno. A implementação deve ser feita em Matlab, Octave ou Scilab.

c) Apresentar gráfico da curva $V(t)$ para o período de integração.

d) Apresentar gráfico da curva $V'(t)$ para o período de integração.

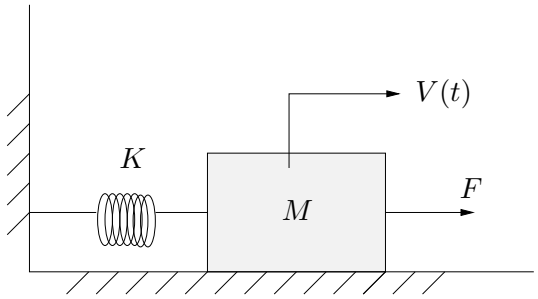


Figura 8.13: Sistema mecânico

Exercício 8.5 A dinâmica do movimento do sistema de dois pêndulos descrito Figura 8.14 pode ser descrita pelas equações diferenciais abaixo:

$$\begin{cases} \ddot{\theta}_1 &= -d_1^{-1}(d_2\ddot{\theta}_2 + \phi_1) \\ \ddot{\theta}_2 &= \left(m_2\alpha_2^2 + I_2 - \frac{d_2^2}{d_1}\right)^{-1}\left(u + \frac{d_2}{d_1}\phi_1 - m_2 l_1 \alpha_2 \dot{\theta}_1^2 \sin\theta_2 - \phi_2\right) \\ d_1 &= m_1\alpha_1^2 + m_2(l_1^2 + \alpha_2^2 + 2l_1\alpha_2\cos\theta_2) + I_1 + I_2 \\ d_2 &= m_2(\alpha_2^2 + l_1\alpha_2\cos\theta_2) + I_2 \\ \phi_1 &= -m_2 l_1 \alpha_2 \dot{\theta_2}^2 \sin\theta_2 - 2m_2 l_1 \alpha_2 \dot{\theta}_2\dot{\theta}_1 \sin\theta_2 + (m_1\alpha_1 + m_2 l_1) g\cos(\theta_1 - \tfrac{\pi}{2}) + \phi_2 \\ \phi_2 &= m_2\alpha_2 g\cos(\theta_1 + \theta_2 - \tfrac{\pi}{2}) \end{cases} \tag{8.82}$$

onde u é o torque aplicado na junta, $m_1 = m_2 = 1$ são as massas das hastes, $l_1 = l_2 = 1$ são os comprimentos das hastes, $\alpha_1 = \alpha_2 = 1/2$ são os comprimentos até o centro de massa das barras, $I_1 = I_2 = 1$ são os momentos de inércia das barras, e $g = 9.8$ é a aceleração da gravidade. Todas as unidades estão no sistema internacional.

a) Fazendo $x_1 = \theta_1, x_2 = \dot{\theta}_1, x_3 = \theta_2, x_4 = \dot{\theta}_2$ e $x = (x_1, x_2, x_3, x_4)$, obter um sistema de equações diferenciais

$$\dot{x} = F(x, u)$$

equivalente ao sistema de equações (8.82).

b) Quando nenhum torque externo é aplicado, $u(t) = 0$, obter a trajetória de $x(t)$ para $t \in [0, 10s]$ usando o método de Runge-Kutta de 2a ordem e fazendo $x(0) = [\frac{\pi}{7}, 0.2, \frac{\pi}{10}, -0.1]$. Apresentar em gráficos as curvas $x_1(t), \ldots, x_4(t)$. Qual é o valor de $x(10)$?

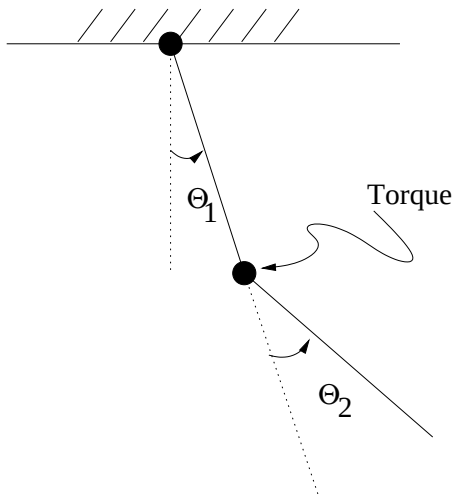


Figura 8.14: Sistema mecânico de dois pêndulos

Exercício 8.6 Seja o PVI: $\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \quad , \quad t_0 \leqslant t \leqslant t_f \end{cases}$

a) Represente graficamente e dê uma explicação para um passo do método de Euler.

b) Elabore um *pseudo-código* para a implementação computacional do método de Runge-Kutta de 2^a-ordem. Considere:

i) dados de entrada (fornecidos pelo usuário): a condição inicial, o intervalo de cálculo e o número de iterações desejadas;

ii) que para valores particulares de (t, y), o valor numérico de $f(t, y(t))$ obtidos pela chamada de uma "rotina" (ou função auxiliar) com a sintaxe: `Fty = avaliaF(t,y);`

iii) dados de saída: o programa correspondente ao pseudo-código dever fornecer os valores aproximados de $y(t)$ ao longo das iterações e os valores correspondentes da variável independente.

Referências Bibliográficas

[1] D. M. Cláudio and J. M. Marins. *Cálculo Numérico Computacional*. Atlas, 1994.

[2] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[3] C. C. Dyer and S. S. Ip. An Elementary Introduction to Scientific Computing. Division of Physical Sciences, Univerisity of Toronto at Scarborough, January 2000.

[4] H. Jeffreys and B. S. Jeffreys. *Methods of Mathematical Physics*. Cambridge University Press, Cambridge, England, 3rd edition, 1988.

[5] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England, 2nd edition, 1992.

[6] S. L. Salas, E. Hille, and J. T. Anderson. *Calculus: One and Several Variables, with Analytic Geometry*. John Wiley and Sons, New York, NY, 5th edition, 1986.

[7] L. Vandenberghe. Applied Numerical Computing. Lecture Notes for EE103, 2001.

Apêndice A

Fundamentos Matemáticos

A.1 Limites e Continuidade

Informalmente,

$$\lim_{x \to c} f(x) = l$$

significa que para x próximo mas diferente de c, $f(x)$ está próximo de l, o que significa dizer que se

$|x - c|$ é pequeno mas diferente de zero, então $|f(x) - l|$ é pequeno.

Dizer que $\lim_{x \to c} f(x) = l$ significa dizer que $|f(x) - l|$ pode tornar-se arbitrariamente pequeno simplesmente fazendo $|x - c|$ suficientemente pequeno mas diferente de zero. Se você tomar $\epsilon > 0$, então $|f(x) - l|$ pode se tornar menor do que ϵ se $0 < |x - c| < \delta$ para um δ suficientemente pequeno. A definição a seguir formaliza este princípio.

Definição A.1 *O limite de uma função*

$$\lim_{x \to c} f(x) = l \text{ sse } \begin{cases} \text{para cada } \epsilon > 0 \text{ existe } \delta > 0 \text{ tal que} \\ \text{se } 0 < |x - c| < \delta, \text{ então } |f(x) - l| < \epsilon \end{cases}$$

Exemplo A.1 *Mostre que* $\lim_{x \to 2}(2x - 1) = 3$.

(*Encontrando* δ) Seja $\epsilon > 0$. Desejamos encontrar $\delta > 0$ tal que,

$$\text{se } 0 < |x - 2| < \delta, \text{ então } |(2x - 1) - 3| < \epsilon$$

Primeiramente, estabelecemos a conexão entre

$$|(2x - 1) - 3| \text{ e } |x - 2|.$$

A conexão é simples,

$$|(2x-1)-3| = |2x-4| = 2|x-2|. \tag{A.1}$$

Para fazer $|(2x-1)-3|$ menor do que ϵ, precisamos apenas fazer $|x-2|$ ser duas vezes menor. Isso sugere a escolha de $\delta = \frac{\epsilon}{2}$. (*Mostrando que funciona*) Para mostrar que esta escolha funciona, note que se $0 < |x-2| < \frac{\epsilon}{2}$, então $2|x-2| < \epsilon$ e de (A.1) temos que $|(2x-1)-3| < \epsilon$.

Definição A.2 *(O limite à esquerda de uma função) Seja f uma função definida no intervalo (a, c).*

$$\lim_{x \to c^-} f(x) = l \text{ sse } \begin{cases} \text{para cada } \epsilon > 0 \text{ existe } \delta > 0 \text{ tal que} \\ \text{se } c - \delta < x < c, \text{ então } |f(x) - l| < \epsilon \end{cases}$$

Definição A.3 *(O limite à direita de uma função) Seja f uma função definida no intervalo (c, d).*

$$\lim_{x \to c^+} f(x) = l \text{ sse } \begin{cases} \text{para cada } \epsilon > 0 \text{ existe } \delta > 0 \text{ tal que} \\ \text{se } c < x < c + \delta, \text{ então } |f(x) - l| < \epsilon \end{cases}$$

Segue das definições acima que

$$\lim_{x \to c} f(x) = l \text{ sse } \lim_{x \to c^+} = l \text{ e } \lim_{x \to c^-} = l$$

Dizemos que um processo é contínuo se ele é realizado sem interrupção e sem mudanças bruscas. Na matemática, a palavra "contínuo" tem um significado diferente.

Seja f uma função definida no intervalo aberto $(c-\delta, c+\delta)$.

Definição A.4 *A função f é contínua no ponto c se e somente se $\lim_{x \to c} f(x) = c$.*

Se o domínio da função f contém o intervalo aberto $(c-\delta, c+\delta)$, então há apenas duas razões para f não ser contínua no ponto c: i) $f(x)$ não tem um limite quando x tende para c, neste caso dizemos que c é uma descontinuidade; ii) $f(x)$ tem um limite l, mas o limite l é diferente de c. No último caso, podemos eliminar a descontinuidade redefinindo f de forma que $f(x) = l$. A Figura A.1 ilustra a transformação de uma função descontínua (a) em uma função contínua (b).

Uma função que é contínua dentro de um intervalo não tem saltos nem interrupções, portanto a sua curva não é quebrada. Com base nestes princípios, dois teoremas podem ser estabelecidos.

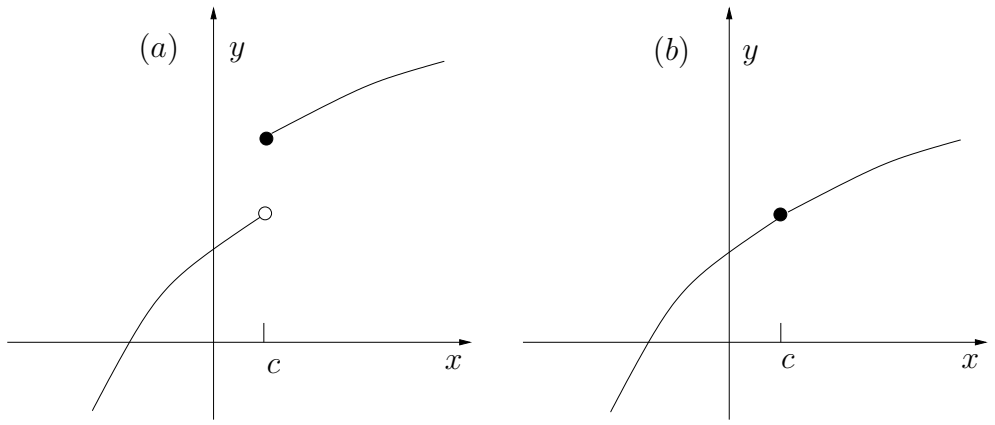


Figura A.1: Eliminando a descontinuidade de uma função.

Teorema A.1 *(Teorema do valor intermediário) Se f é contínua no intervalo $[a, b]$ e C é um número entre $f(a)$ e $f(b)$, então existe $c \in [a, b]$ tal que $f(c) = C$.*

Teorema A.2 *(Teorema do máximo e mínimo) Se f é contínua em $[a, b]$, então f tem um valor máximo M e mínimo m no intervalo $[a, b]$.*

A.2 Diferenciação

Considere uma função $f : \mathbb{R} \to \mathbb{R}$ e o ponto $(x, f(x))$ de seu gráfico, conforme Figura A.2. Como que se obtém, quando esta existe, a tangente da curva no ponto $(x, f(x))$?

Para responder esta questão, tomamos um número pequeno $h \neq 0$ e na curva obtemos o ponto $((x + h), f(x + h))$, conforme ilustração na Figura A.3. Agora traçamos a reta secante que passa pelos pontos $(x, f(x))$ e $((x + h), f(x + h))$. Dependendo de $h > 0$ ou $h < 0$, a secante terá diferentes inclinações. À medida que h tende para zero da direita, a secante tende para uma posição limite, o mesmo ocorrendo quando h tende para zero da esquerda. A reta nesta posição limite é dita "tangente da curva no ponto $(x, f(x))$".

Uma vez que as secantes tem inclinação dada por:

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

espera-se que a tangente, a posição limite destas secantes, tenha inclinação

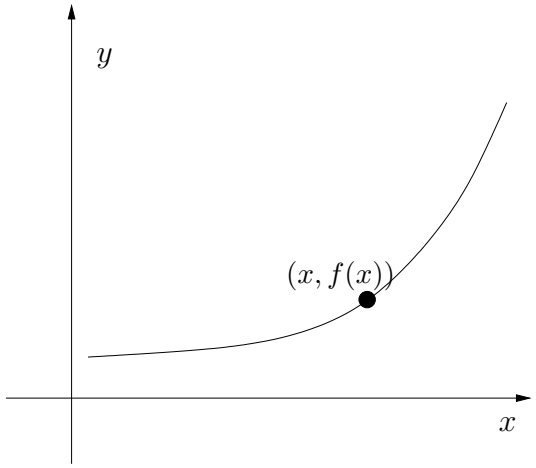


Figura A.2: Função exemplo.

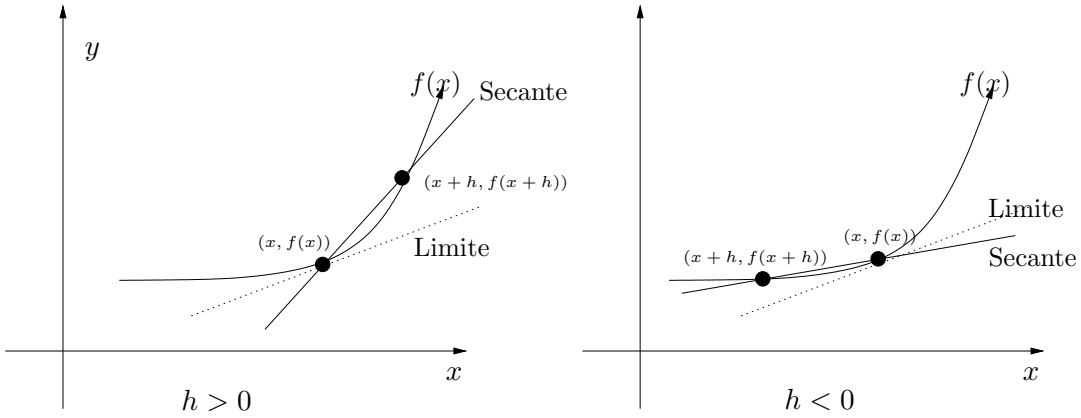


Figura A.3: Secantes da função.

dada por:

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}.$$

Definição A.5 *Uma função f é dita diferenciável no ponto x se*

$$\lim_{x\to c} \frac{f(x+h)-f(x)}{h} \text{ existe.}$$

Se este limite existe, então a derivada de f no ponto x é denotada por $f'(x)$.

Teorema A.3 *Se f é diferenciável no ponto x, então f é contínua em x.*

Teorema A.4 *(Regra do produto) Se f e g são duas funções diferenciáveis no ponto x, então*

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$

Definição A.6 *(Derivada da soma e múltiplo escalar) Seja α um número real. Se f e g são diferenciáveis no ponto x, então $f+g$ and αf são diferenciáveis no ponto x tal que:*

$$(f+g)'(x) = f'(x) + g'(x) \text{ and } (\alpha f)'(x) = \alpha f'(x).$$

Aqui vamos considerar a diferenciação de funções compostas. Suponha que y é uma função diferenciável em u:

$$y = f(u)$$

e u é uma função diferenciável em x:

$$u = g(x).$$

Portanto, y é uma função de x:

$$y = f(u) = f(g(x)).$$

Podemos dizer que y tem uma derivada em relação a x? Sim, y tem uma uma derivada em relação a x dada pela fórmula:

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

A fórmula acima é conhecida como *regra da cadeia*, a qual diz que a taxa de variação de y com respeito a x é a taxa de variação de y com respeito a u

vezes a taxa de variação de u com respeito a x. Como exemplo, considere as funções a seguir:

$$y = 2u \text{ e } u = 3x.$$

Então, temos que

$$y = 2x.$$

Note que

$$\frac{dy}{dx} = 6, \quad \frac{dy}{du} = 2, \quad \frac{du}{dx} = 3$$

o que nos leva a confirmar a regra da cadeia pois:

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

Teorema A.5 *(Regra da cadeia) Se g é diferenciável no ponto x e f é diferenciável no ponto $g(x)$, então a composição $f.g$ é diferenciável no ponto x tal que*

$$(f.g)'(x) = f'(g(x))g'(x).$$

A.2.1 Aplicação de Diferenciação

Um tanque cujo corte facial é um triângulo eqüilátero recebe água na taxa de 4 cm^3/min. Tendo a abertura do tanque 12 cm de comprimento, com que velocidade a água está subindo no instante em que ela atinge a profundidade de $1\frac{1}{2}$ cm?

A Figura A.4 ilustra o corte facial do tanque. Seja x a profundidade da água e V o volume de água em cm^3. Sabemos que $\frac{dV}{dt} = 4$ cm^3/min e desejamos saber $\frac{dx}{dt}$ quando $x = \frac{3}{2}$ cm. Note que $l = \frac{x}{\tan(60^o)} = \frac{x\sqrt{3}}{3}$. Portanto, a área da seção facial é $lx = \frac{\sqrt{3}}{3}x^2$ e o volume de água armazenado no tanque é $12\left(\frac{x^2\sqrt{3}}{3}\right) = 4\sqrt{3}x^2$.

Diferenciando $V = 4\sqrt{3}x^2$ com respeito a t obtemos

$$\frac{dV}{dt} = 8\sqrt{3}x\frac{dx}{dt}.$$

Substituindo $x = \frac{3}{2}$ e $\frac{dV}{dt} = 4$, temos

$$4 = 8\sqrt{3}\left(\frac{3}{2}\right)\frac{dx}{dt}$$

o que nos leva a concluir que

$$\frac{dx}{dt} = \frac{1}{9}\sqrt{3}.$$

No instante em que a água atinge a altura $1\frac{1}{2}cm$, o nível da água está subindo a uma taxa de $\frac{1}{9}\sqrt{3}cm/s$.

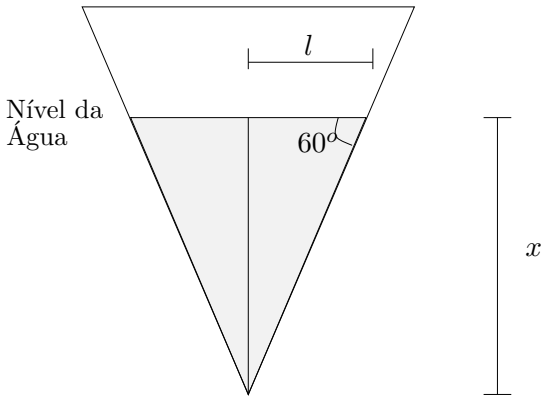


Figura A.4: Corte facial do tanque.

A.3 Teorema do Valor Médio

O Teorema do Valor-Médio foi primeiramente enunciado pelo matemático Joseph Louis Lagrange (1736–1813), cujas aplicações podem ser encontradas em vários problemas da matemática.

Teorema A.6 *(*Torema do Valor Médio*) Se f é uma função diferenciável em (a, b) e contínua em $[a, b]$, então existe um número $c \in (a, b)$ tal que*

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

O número c é a inclinação da reta l que passa através dos pontos $(a, f(a))$ e $(b, f(b))$. Dizer que existe pelo menos um número c tal que

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

significa dizer que o gráfico da função f tem pelo menos um ponto $(c, f(c))$ no qual a tangente é paralela a l. Esta configuração é ilustrada na Figura A.5.

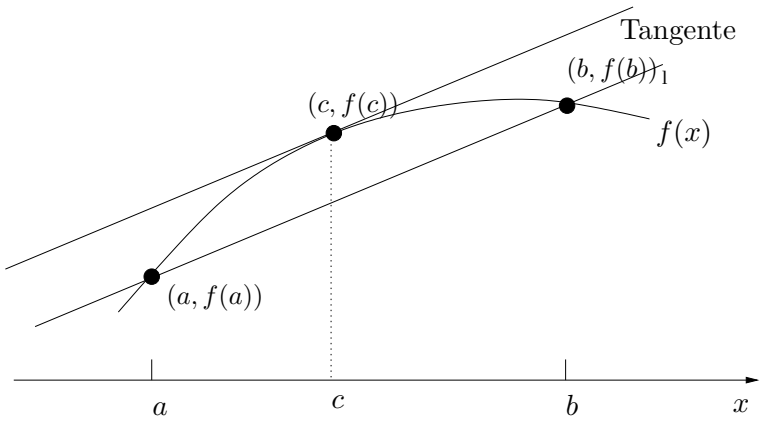


Figura A.5: Ilustração do Teorema do Valor Médio.

A.4 Máximos e Mínimos

Em problemas de Engenharia e Física, rotineiramente se deseja determinar quão grande ou quão pequena uma certa quantidade pode ser. Se o problema admite uma formulação matemática, então podemos em princípio reduzi-lo ao problema de encontrar o valor mínimo ou máximo de uma certa função. No que segue, vamos considerar os valores máximos e mínimos de uma função em um intervalo aberto.

Definição A.7 (Mínimo/Máximo Local) *Uma função f tem um* máximo local no ponto c *sse*

$$f(c) \geq f(x) \text{ para todo } x \text{ suficiente próximo de } c.$$

A função tem um mínimo local no ponto c *sse*

$$f(c) \leq f(x) \text{ para todo } x \text{ suficiente próximo de } c.$$

As noções de pontos de máximo e mínimo locais são ilustradas na Figura A.6.

A.5 Introdução a Equações Diferenciais

Na Natureza é comum encontrarmos sistemas cujas grandezas variam no tempo. Se uma quantitade y varia no tempo de acordo com uma equação diferencial, então y é tipicamente descrita por uma equação que envolve y e

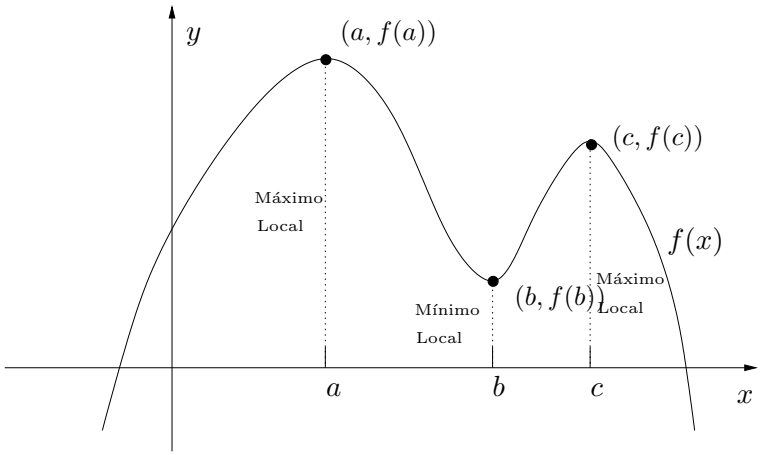


Figura A.6: Pontos de máximo e mínimo de uma função.

suas derivadas. Tal equação é conhecida como *equação diferencial*. Exemplos de equações diferenciais são:

$$y' = 2y + e^{-y^2},\ 2y'' - y' + 2y = 0,\ y'' + 2y' - y = x.$$

Equações diferenciais são essencias na modelagem matemática de fenômenos físicos, tendo inúmeras aplicações reais em engenharia, física e matemática.

Muitas vezes os problemas de interesse são tão complexos que a análise destes torna-se intratável. A investigação do impacto de novas políticas cambiais na economia de um país e a influência do acúmulo de resíduos poluentes em um ecosistema configuram sistemas de complexidade considerável. Em tais sistemas, o tópico de interesse pode ser identificado com certa facilidade mas, por outro lado, é muito difícil de se chegar a conclusões. As relações entre as grandezas podem não ser aparentes. Em tais situações, se faz necessário o emprego de modelos simplificados que tomam como hipóteses premissas restritivas. Por exemplo, na modelagem de sistemas mecânicos podemos desconsiderar as variações da gravidade e a resistência do ar para velocidades baixas. Contudo, é importante que as hipóteses sejam enunciadas de forma a deixar claro as limitações e a região de validade. Se as hipóteses podem ser escritas em notação matemática, então as ferramentas de modelagem e solução matemática podem ser empregadas no tratamento do problema. O processo de transcrever, formular, analisar e resolver um problema em um contexto matemático é conhecido como *modelagem matemática*.

O objetivo da modelagem matemática é melhor enteder um fenômeno do mundo real, o que não é uma tarefa trivial. Muitas vezes, desejamos construir um modelo que nos permita fazer predições que por sua vez podem ser empregadas para influenciar a evolução de eventos. Por exemplo, o modelo de

um tanque de biomassa pode ser utilizado na determinação da temperatura e composição que garantam maior geração de energia.

Equações Diferenciais de Primeira Ordem

A ordem de uma equação diferencial é a ordem da derivada mais alta que aparece na equação. As equações

$$y' + 2y = x^3 \text{ and } y' - xy/2 = e^x$$

são equações de primeira ordem;

$$y'' - 2y' + y = 2x \text{ and } y'' + 4y = 0$$

são equações de segunda ordem. Uma função y é dita uma *solução* da equação diferencial se ela satisfaz a equação. Por exemplo, a função

$$y = x + e^{-2x}$$

é uma solução da equação diferencial

$$y' + 2y = 2x + 1.$$

Basta verificar que

$$y' + 2y = (x + e^{-2x})' + 2(x + e^{-2x}) = 1 - 2e^{-2x} + 2x + 2e^{-2x} = 2x + 1.$$

Para ilustrar a solução de equações diferenciais de primeira ordem, vamos considerar equações lineares na forma

$$y' + p(x)y = q(x)$$

onde p e $q(x)$ são funções contínuas. No caso mais simples, quando $p(x) = 0$ para todo x, a equação se reduz a

$$y' = q(x). \tag{A.2}$$

As soluções desta equação são antiderivadas de q podendo ser escritas na forma $y = Q(x)$. Portanto,

$$y = Q(x) + c \tag{A.3}$$

é uma solução com c sendo uma constante arbitrária. A equaçào (A.3) é dita *solução geral* e qualquer solução de (A.2) pode ser obtida ajustando o valor da constante c. A função $y = x^2$, por exemplo, resolve a equação diferencial $y' = 2x$. A solução geral é portanto $y = x^2 + c$.

Para resolver uma equação na forma

$$y' + p(x)y = q(x)$$

primeiramente calculamos

$$P(x) = \int p(x)dx$$

sem considerar constantes. Multiplicamos a equação diferencial por $e^{P(x)}$ obtendo

$$e^{P(x)}y' + e^{P(x)}p(x)y = e^{P(x)}q(x).$$

O lado esquerdo da equação acima é

$$\frac{d}{dx}[e^{P(x)}y]$$

portanto, a equação pode ser escrita como

$$\frac{d}{dx}[e^{P(x)}y] = e^{P(x)}q(x).$$

Integrando, obtemos

$$e^{P(x)}y = \int e^{P(x)}q(x)dx + c$$

e a solução é dada por

$$y = e^{-P(x)}\{\int e^{P(x)}q(x)dx + c\}.$$

Modelo Presa-Predador

Objetivando ilustrar a modelagem matemática, vamos considerar um cenário de duas espécies onde uma serve de alimento a outra. Digamos que as espécies são raposas e coelhos. Se a população de coelhos cresce, então as raposas encontram presas com mais facilidade e em decorrência de melhor alimento também crescem em número. Após algum tempo, a população de coelhos decresce. Como conseqüência, a população de raposas também decresce em virtude da dificuldade de encontrar alimento. A redução da população de raposas, por sua vez, facilita o crescimento da população de coelhos, levando de volta ao início do ciclo.

Seja x o número de raposas e y o número de coelhos no instante t. Vamos considerar as espécies separadamente. Se não há raposas, a população de coelhos cresce exponencialmente, o que pode ser modelado pela equação

$$\frac{dy}{dt} = ay, \quad a > 0. \tag{A.4}$$

Por outro lado, se não há coelhos, a população de raposas decresce exponencialmente conforme modelo

$$\frac{dx}{dt} = -bx, \quad b > 0. \tag{A.5}$$

Para considerar a interação entre espécies, vamos assumir que há um número abundante de coelhos e raposas. Vamos assumir também que a taxa na qual raposas matam coelhos é proporcional ao produto xy. Assim, subtraímos um termo proporcional a xy de (A.4), dessa forma modelando a redução da população de coelhos. Adicionamos também um termo proporcional a xy à expressão (A.5), modelando o crescimento da população de raposas. O modelo em equações diferenciais fica

$$\frac{dy}{dt} = ay - sxy, \quad \frac{dx}{dt} = rxy - bx, \text{ onde } \ a, b, r, s > 0.$$

O termo xy é adotado em vez da expressão $x + y$ pois a quantitdade xy é a maior possível quando x e y são aproximadamente iguais. O maior número de presas serão capturadas quando não há um número relativo elevado de raposas a procura de presas e quanto não há um número muito pequeno de coelhos.

As equações diferenciais A.5 forma um sistema de equações. Podemos resolver o sistema eleminando a variável independente t:

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}$$

que pode ser escrita como

$$\frac{dy}{dx} = \frac{y(a - sx)}{x(ry - b)}.$$

Separando as variáveis obtemos

$$\begin{aligned}
\frac{dy}{y(a-sx)} &= \frac{dx}{x(ry-b)}, \\
\frac{ry-b}{y}dy &= \frac{(a-sx)}{x}dx, \\
ry - b\ln y + sx - a\ln x &= c, \\
ry + sx - \ln x^a y^b &= c, \\
ry + sx - c &= \ln x^a y^b, \\
e^{ry+sx-c} &= x^a y^b, \\
e^{-c} &= \frac{x^a y^b}{e^{ry}e^{sx}}.
\end{aligned}$$

Fazendo $k = e^{-c}$, a solução fica

$$k = \frac{x^a y^b}{e^{ry}e^{sx}}. \tag{A.6}$$

Note que k pode ser calculado com base nos valores iniciais de x e y ($x(0)$ e $y(0)$). Observe também que k não depende do tempo. Para analisar o comportamento da solução, note que o lado direito de (A.6) é composto de dois termos similares da forma:

$$f(z) = \frac{z^p}{e^{mz}}, \quad p, m > 0.$$

O gráfico da curva $f(z)$ tem a forma dada na Figura A.7.

Se fixarmos o valor de x em (A.6) e resolvermos a equação para y, obteremos dois valores. Ou seja, $f(z)$ assume este valor em dois pontos, z_1 e z_2. A exceção ocorre quando $y = b/r$ onde $z = p/m$. Nesse valor de y, x assume seu valor máximo ou mínimo. De forma similar, linhas horizontais ordinariamente interceptam o gráfico de (A.6) duas vezes, sendo que y assume seu valor máximo ou mínimo quando x assume o valor a/s. O gráfico de (A.6) é uma curva conhecida como *trajetória*. Trajetórias para dois valores de K e parâmetros fixos para a, b, r e s aparecem na Figura A.8.

Para determinar a direção das trajetórias, conforme indicado pelas setas, basta verificar que

$$\frac{dx}{dy} = x(ry-b) > 0 \text{ sse } y > b/r.$$

Portanto, o movimento é no sentido horário. O ponto $E = (a/s, b/r)$ é um ponto de equilíbrio. As duas populações não mudam com o passar do tempo. Na prática tal equilíbrio não ocorre. O comportamento cíclico ilustrado pela figura melhor representa o caso real.

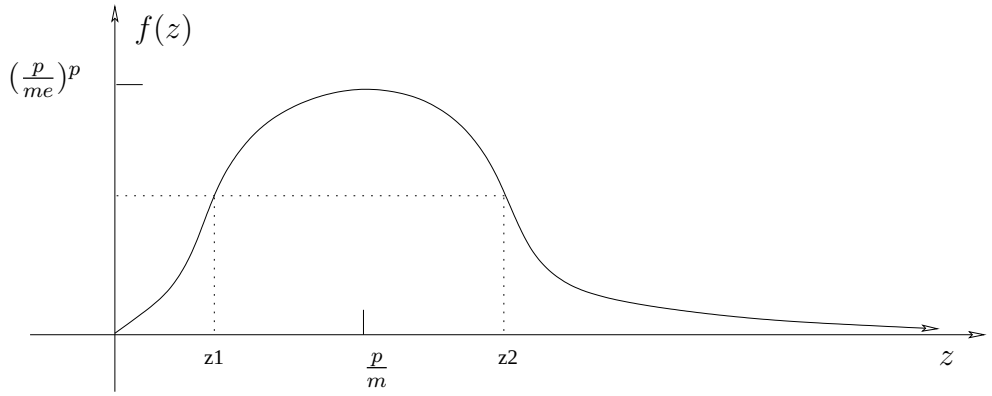


Figura A.7: Gráfico da função $f(z) = \frac{z^p}{e^{mz}},\ p, m > 0$.

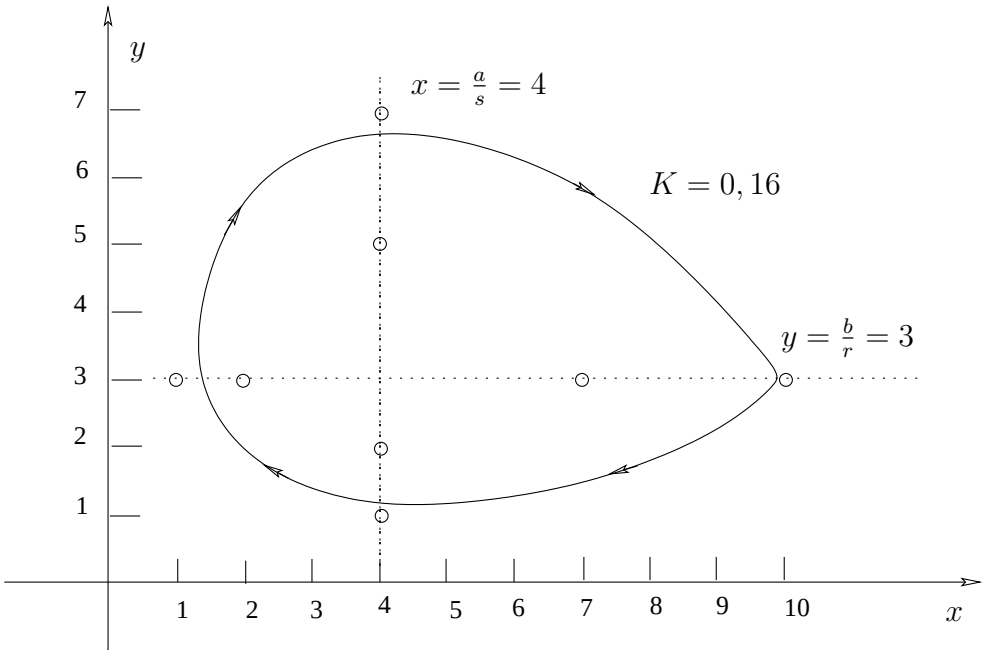


Figura A.8: Trajetórias para $K = 0,16$.

A.6 Vetores

Uma tripla $\mathbf{x} = (x_1, x_2, x_3)$ que define uma direção $\mathbf{v} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$ é um vetor do plano cartesiano, onde $\mathbf{i}$ é o vetor de comprimento unitário que dá a direção do eixo x, $\mathbf{j}$ é o vetor de comprimento unitário que dá a direção do eixo y, e $\mathbf{k}$ é unitário e dá a direção do eixo z. Os vetores $\mathbf{i}$, $\mathbf{j}$ e $\mathbf{k}$ são ortogonais entre si. Note que $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$ e $\mathbf{k} = (0, 0, 1)$. O ponto $\mathbf{0} = (0, 0, 0)$ é a origem do plano cartesiano e serve de referência. A Figura A.9 ilustra o plano cartesiano. Neste apêndice vamos diferenciar escalares de vetores adotando fonte em negrito para os vetores.

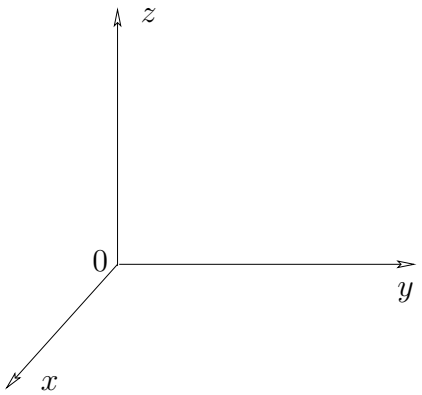


Figura A.9: Plano cartesiano.

A adoção do plano cartesiano para representação de vetores facilita operações básicas sobre vetores. Por exemplo, a soma vetorial de dois vetores $\mathbf{a} = (a_1, a_2, a_3)$ e $\mathbf{b} = (b_1, b_2, b_3)$ pode ser obtida fazendo:

$$\mathbf{a} + \mathbf{b} = (a_1, a_2, a_3) + (b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$$

A multiplicação de um vetor $\mathbf{a}$ por um escalar α também é simples, sendo dada por:

$$\alpha\mathbf{a} = (\alpha a_1, \alpha a_2, \alpha a_3)$$

A soma vetorial de dois vetores $\mathbf{a}$ e $\mathbf{b}$ está exemplificada na Figura A.10.

Dois vetores não nulos $\mathbf{a}$ e $\mathbf{b}$ são ditos *paralelos* se $\mathbf{a} = \alpha\mathbf{b}$ para algum escalar α. Dois vetores paralelos $\mathbf{a}$ e $\mathbf{b}$ têm a *mesma direção* se $\alpha > 0$; eles têm *direções opostas* se $\alpha < 0$.

O comprimento de um vetor é medido por meio de normas. A norma l-2 ou norma Euclidiana de um vetor $\mathbf{x} = (x_1, x_2, x_3)$ é definida como:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

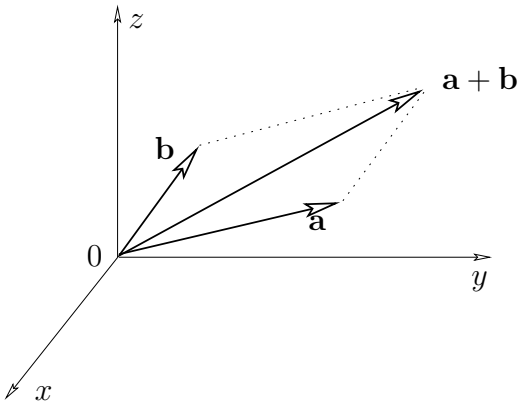


Figura A.10: Soma vetorial.

Note que toda a norma $\|\cdot\|$ satisfaz as propriedades:

(1) $\|\mathbf{x}\| \geq 0$ para todo $\mathbf{x} \in \mathbb{R}^3$;

(2) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ para todo $\mathbf{x} \in \mathbb{R}^3$ e $\alpha \in \mathbb{R}$; e

(3) $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ para todo $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$.

A.6.1 Produto Interno

O produto interno tem inúmeras aplicações em física e geometria. Para dois vetores arbitrários $\mathbf{a} = (a_1, a_2, a_3)$ e $\mathbf{b} = (b_1, b_2, b_3)$, o produto interno é um escalar denotado por $\mathbf{a} \cdot \mathbf{b}$ e definido por:

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3$$

Observe que $\mathbf{a} \cdot \mathbf{a} = \|\mathbf{a}\|^2$. O produto interno de um vetor $\mathbf{a}$ qualquer com o vetor nulo é zero:

$$\mathbf{a} \cdot \mathbf{0} = 0 \text{ e } \mathbf{0} \cdot \mathbf{a} = 0$$

Podemos mostrar que o produto interno é comutativo verificando que:

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3 = b_1a_1 + b_2a_2 + b_3a_3 = \mathbf{b} \cdot \mathbf{a}$$

Escalares também são passíveis de fatoração conforme a dedução abaixo demonstra:

$$\begin{aligned} \alpha\mathbf{a} \cdot \beta\mathbf{b} &= \alpha\beta a_1b_1 + \alpha\beta a_2b_2 + \alpha\beta a_3b_3 \\ &= \alpha\beta(a_1b_1 + a_2b_2 + a_3b_3) \\ &= \alpha\beta\mathbf{a} \cdot \mathbf{b} \end{aligned}$$

Da maneira semelhante ao desenvolvido acima pode-se mostrar que o produto interno é distributivo, ou seja, $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$.

Se dois vetores $\mathbf{a}$ e $\mathbf{b}$ são não nulos, então o produto interno $\mathbf{a} \cdot \mathbf{b}$ pode ser interpretado geometricamente tomando como base o triângulo formado pelos vetores $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{a} - \mathbf{b}$ conforme indicado na Figura A.11. Pela lei dos cosenos, temos que:

$$\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\| \cos\theta$$

o que nos leva a:

$$\begin{aligned}
2\|\mathbf{a}\|\|\mathbf{b}\| \cos\theta &= \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2 \\
&= [a_1^2 + a_2^2 + a_3^2] + [b_1^2 + b_2^2 + b_3^2] \\
&\quad -[(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2] \\
&= [a_1^2 + a_2^2 + a_3^2] + [b_1^2 + b_2^2 + b_3^2] \\
&\quad -[a_1^2 - 2a_1b_1 + b_1^2 + a_2^2 - 2a_2b_2 + b_2^2 + a_3^2 - 2a_3b_3 + b_3^2] \\
&= 2a_1b_1 + 2a_2b_2 + 2a_3b_3 \\
&= 2\mathbf{a} \cdot \mathbf{b}
\end{aligned}$$

portanto, concluímos que:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta \tag{A.7}$$

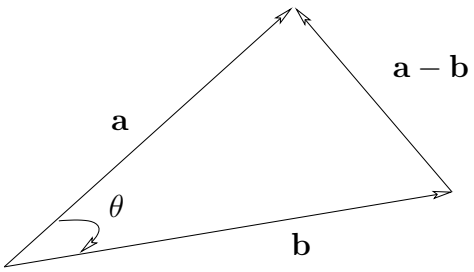


Figura A.11: Produto interno.

A partir da expressão (A.7) podemos verificar que o produto interno nos dá uma medida do quanto as direções de dois vetores se assemelham. À medida que as direções de dois vetores se tornam opostas o produto interno entre eles decresce.

1) Quanto dois vetores $\mathbf{a}$ e $\mathbf{b}$ têm a mesma direção, $\theta = 0$, o produto interno se torna:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\|$$

que é o máximo valor possível para o produto interno $\mathbf{a} \cdot \mathbf{b}$.

2) Quando $\mathbf{a}$ e $\mathbf{b}$ têm direções opostas, $\theta = -\pi$, o produto interno assume o valor:

$$\mathbf{a} \cdot \mathbf{b} = -\|\mathbf{a}\|\|\mathbf{b}\|$$

que é o mínimo valor possível para $\mathbf{a} \cdot \mathbf{b}$.

3) Quando $\mathbf{a}$ e $\mathbf{b}$ são ortogonais, $\theta = \frac{\pi}{2}$, o produto interno assume o valor:

$$\mathbf{a} \cdot \mathbf{b} = 0$$

Concluímos que dois vetores são perpendiculares se e somente se o produto interno entre eles é nulo.

A.6.2 Projeções

Se $\mathbf{b}$ é um vetor não nulo, então um vetor $\mathbf{a}$ arbitrário pode ser escrito de maneira única como a soma de um vetor $\mathbf{b}^{\parallel}$ paralelo a $\mathbf{b}$ e um vetor $\mathbf{b}^{\perp}$ ortogonal a $\mathbf{b}$:

$$\mathbf{a} = \mathbf{b}^{\parallel} + \mathbf{b}^{\perp}$$

Uma ilustração da decomposição de $\mathbf{a}$ em $\mathbf{b}^{\parallel}$ (projeção de $\mathbf{a}$ sobre $\mathbf{b}$) e $\mathbf{b}^{\perp}$ é dada na Figura A.12.

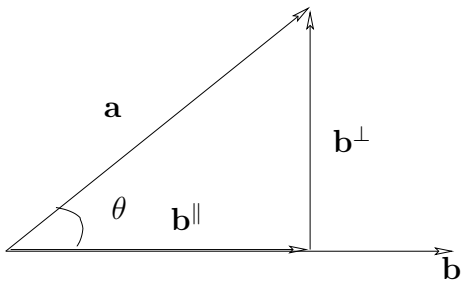


Figura A.12: Produto interno.

A projeção de $\mathbf{a}$ sobre $\mathbf{b}$ é denotada por $\mathbf{proj}_{\mathbf{b}}(\mathbf{a})$. Para a ilustração da Figura A.12, $\mathbf{proj}_{\mathbf{b}}(\mathbf{a}) = \mathbf{b}^{\parallel}$. Note que a projeção de $\mathbf{a}$ em $\mathbf{b}$ é um múltiplo α do vetor $\mathbf{b}$, ou seja:

$$\mathbf{proj}_{\mathbf{b}}(\mathbf{a}) = \mathbf{b}^{\parallel} = \alpha\mathbf{b}$$

Podemos calcular α por meio da relação trigonométrica dada na Figura A.12:

$$\begin{aligned}\alpha &= \|\mathbf{a}\|\cos\theta \\ &= \|\mathbf{a}\|\frac{\|\mathbf{b}\|}{\|\mathbf{b}\|}\cos\theta \\ &= \frac{1}{\|\mathbf{b}\|}\mathbf{a}\cdot\mathbf{b}\end{aligned}$$

Logo, concluímos que a projeção de $\mathbf{a}$ em $\mathbf{b}$ pode ser definida como:

$$\begin{aligned}\mathbf{proj_b}(\mathbf{a}) &= \alpha\mathbf{b} \\ &= \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{b}\|}\mathbf{b} \\ &= (\mathbf{a}\cdot\mathbf{b})\mathbf{u_b}\end{aligned}$$

onde $\mathbf{u_b} = \mathbf{b}/\|\mathbf{b}\|$ é o vetor unitário na direção de $\mathbf{b}$. Vale ressaltar que a projeção de um vetor sobre outro pode ser facilmente calculada a partir do produto interno entre eles. Para calcular a representação do vetor $\mathbf{a}$ em termos de componentes ortogonais e paralelas ao vetor $\mathbf{b}$, basta calcular $\mathbf{b}^{\|}$ e depois subtrair de $\mathbf{a}$ para obter $\mathbf{b}^{\perp}$.

A desigualdade de *Schwartz*, $|\mathbf{a}\cdot\mathbf{b}| \leq \|\mathbf{a}\|\|\mathbf{b}\|$, é facilmente demonstrada a partir do produto interno:

$$\begin{aligned}|\mathbf{a}\cdot\mathbf{b}| &= |\|\mathbf{a}\|\|\mathbf{b}\|\cos\theta| \\ &= \|\mathbf{a}\|\|\mathbf{b}\|\ |\cos\theta| \\ &\leq \|\mathbf{a}\|\|\mathbf{b}\| \qquad \text{(A.8)}\end{aligned}$$

Fazendo uso da desigualdade de *Schwartz* podemos também demonstrar a *desigualdade triangular*:

$$\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$$

Para tanto, basta verificar que:

$$\begin{aligned}\|\mathbf{a}+\mathbf{b}\|^2 &= (\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) \\ &= \mathbf{a}\cdot\mathbf{a}+\mathbf{a}\cdot\mathbf{b}+\mathbf{b}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b} \\ &= \mathbf{a}\cdot\mathbf{a}+2\mathbf{a}\cdot\mathbf{b}+\mathbf{b}\cdot\mathbf{b} \\ &\leq \mathbf{a}\cdot\mathbf{a}+2|\mathbf{a}\cdot\mathbf{b}|+\mathbf{b}\cdot\mathbf{b} \\ &\leq \|\mathbf{a}\|^2+2\|\mathbf{a}\|\|\mathbf{b}\|+\|\mathbf{b}\|^2 \\ &= (\|\mathbf{a}\|+\|\mathbf{b}\|)^2\end{aligned}$$

Logo, deduzimos que $\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$.

A.6.3 Produto Cruzado

Na mecânica e eletromagnetismo a noção de produto cruzado desempenha um papel relevante. Na Mecânica, o produto cruzado está relacionado à momento de angular, torque e fenômenos de rotação. No Eletromagnetismo, o produto cruzado permite expressas leis das forças que atuam sobre cargas elétricas em movimento. Para vetores

$$\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k} \text{ e } \mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k},$$

o *produto cruzado* $\mathbf{a} \times \mathbf{b}$ é definido como:

$$\mathbf{a} \times \mathbf{b} = (a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}$$

Diferentemente do produto interno $\mathbf{a} \cdot \mathbf{b}$ que é um escalar, o produto interno $\mathbf{a} \times \mathbf{b}$ é um vetor. Tomando como base o determinando de matrizes, podemos expressar o produto cruzado como um determinante:

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix}\mathbf{i} - \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix}\mathbf{j} + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}\mathbf{k}$$

As seguintes propriedade de produto cruzado podem ser verificadas por meio de manipulações algébricas:

1) (*Anticomutatividade*) $\mathbf{b} \times \mathbf{a} = -\mathbf{a} \times \mathbf{b}$;

2) (*Autocancelamento*) $\mathbf{a} \times \mathbf{a} = 0$;

3) (*Fatoração de escalares*) $\alpha\mathbf{a} \times \beta\mathbf{b} = \alpha\beta\mathbf{a} \times \mathbf{b}$;

4) (*Distributividade sobre soma*) $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c})$;

5) (*Ortogonalidade*) $(\mathbf{a} \times \mathbf{b}) \perp \mathbf{a}$ e $(\mathbf{a} \times \mathbf{b}) \perp \mathbf{b}$; e

6) $\|\mathbf{a} \times \mathbf{b}\|^2 = \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2$.

Se um dos vetores $\mathbf{a}$ e $\mathbf{b}$ é nulo, $\mathbf{a} = \mathbf{0}$ ou $\mathbf{b} = \mathbf{0}$, então $\mathbf{a} \times \mathbf{b} = \mathbf{0}$. Caso nenhum dos vetores seja nulo, então podemos expressar o produto interno como:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta$$

e usando a propriedade (6) podemos verificar que:

$$\begin{aligned} \|\mathbf{a} \times \mathbf{b}\|^2 &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\ &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2\|\mathbf{b}\|^2 \cos^2\theta \\ &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2(1 - \cos^2\theta) \\ &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2 \sin^2\theta \end{aligned}$$

que por sua vez implica em:

$$\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\| \sin\theta$$

Esta expressão nos diz que a norma do produto cruzado de dois vetores não nulos é precisamente a área do paralelogramo cujos lados são definidos por estes vetores. Esta propriedade é ilustrada na Figura A.13.

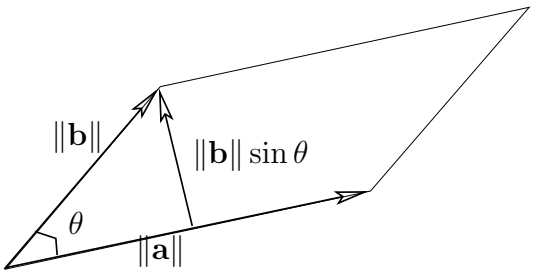


Figura A.13: A área do paralelogramo com lados $\mathbf{a}$ e $\mathbf{b}$ é precisamente $\|\mathbf{a}\times\mathbf{b}\|$.

A.7 Cálculo Vetorial

Dadas funções reais f_1, f_2 e f_3 ($f_j : \mathbb{R} \to \mathbb{R}$), então para cada $t \in \mathbb{R}$ podemos definir o vetor:

$$\mathbf{f}(t) = f_1(t)\mathbf{i} + f_2(t)\mathbf{j} + f_3(t)\mathbf{k} \tag{A.9}$$

que induz uma função vetorial $\mathbf{f}$. Por exemplo, fazendo $f_1(t) = \sin t$, $f_2(t) = \cos t$, $f_3(t) = 0$, obtemos a função vetorial:

$$\mathbf{f}(t) = \sin t\mathbf{i} + \cos t\mathbf{j}$$

Note que para todo t,

$$\|\mathbf{f}(t)\| = \sqrt{\sin^2 t + \cos^2 t} = 1$$

Definição A.8 *(*Limite*) Uma função f dada por (A.9) possui um "limite" em t_0 se e somente se f_1, f_2 e f_3 possuem limite em t_0. Seja $u = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$. Dizemos que:*

$$\lim_{t\to t_0} \mathbf{f}(t) = \mathbf{u} \;\Leftrightarrow\; \begin{array}{l} \lim_{t\to t_0} f_1(t) = x_1, \\ \lim_{t\to t_0} f_2(t) = x_2, \ e \\ \lim_{t\to t_0} f_3(t) = x_3 \end{array} \tag{A.10}$$

Definição A.9 *(*Continuidade*) Dizemos que uma função vetorial* $\mathbf{f}$ *é contínua em* t_0 *se e somente se cada componente é contínua em* t_0*. Matematicamente,* $\mathbf{f}$ *é contínua se e somente se*

$$\lim_{t\to t_0} f_1(t) = f_1(t_0),\ \lim_{t\to t_0} f_2(t) = f_2(t_0)\ e\ \lim_{t\to t_0} f_3(t) = f_3(t_0)$$

De forma mais compacta, $\mathbf{f}$ é contínua $\Leftrightarrow \lim_{t\to t_0} \mathbf{f}(t) = \mathbf{f}(t_0)$.

Definição A.10 *(*Diferenciação*) Uma função vetorial* $\mathbf{f}$ *é diferenciável em* t *se e somente se cada um de seus componentes é diferenciável em* t*. Logo*

$$\mathbf{f}'(t) = f_1'(t)\mathbf{i} + f_2'(t)\mathbf{j} + f_3'(t)\mathbf{k}$$

sendo $\mathbf{f}'(t)$ *dita derivada de* $\mathbf{f}$ *dem* t*.*

Podemos também definir $\mathbf{f}'(t)$ como o limite do vetor do quociente de diferença:

$$\mathbf{f}'(t) = \lim_{h\to 0} \frac{\mathbf{f}(t+h) - \mathbf{f}(t)}{h}$$

Regras de diferenciação:

i. $(\mathbf{f} + \mathbf{g})'(t) = \mathbf{f}'(t) + \mathbf{g}'(t)$

ii. $(\alpha\mathbf{f})'(t) = \alpha\mathbf{f}'(t)$ (para todo escalar α)

iii. $[u(t)\mathbf{f}(t)]' = u'(t)\mathbf{f}(t) + u(t)\mathbf{f}'(t)$

iv. $(\mathbf{f}\cdot\mathbf{g})'(t) = \mathbf{f}(t)\cdot\mathbf{g}'(t) + \mathbf{f}'(t)\cdot\mathbf{g}(t)$

v. $(\mathbf{f}\times\mathbf{g})'(t) = \mathbf{f}'(t)\times\mathbf{g}(t) + \mathbf{f}(t)\times\mathbf{g}'(t)$

vi. $\mathbf{f}(u(t))' = \mathbf{f}'(u(t))u'(t)$ (regra da cadeia)

Definição A.11 *(*Integração*) Para uma função vetorial* $\mathbf{f}$ *contínua em* $[a, b]$*, a integral de* $\mathbf{f}$ *é definida componente a componente:*

$$\int_a^b \mathbf{f}(t)dt = \left(\int_a^b f_1(t)dt\right)\mathbf{i} + \left(\int_a^b f_2(t)dt\right)\mathbf{j} + \left(\int_a^b f_3(t)dt\right)\mathbf{k}$$

Regras de integração:

i. $\int_a^b [\mathbf{f}(t) + \mathbf{g}(t)]dt = \int_a^b \mathbf{f}(t)dt + \int_a^b \mathbf{g}(t)dt$

ii. $\int_a^b \alpha\mathbf{f}(t)dt = \alpha\int_a^b \mathbf{f}(t)dt$ (para todo escalar α)

iii. $\int_a^b \mathbf{c}\cdot\mathbf{f}(t)dt = \mathbf{c}\cdot\int_a^b \mathbf{f}(t)dt$ (para todo vetor $\mathbf{c}$)

iv. $\int_a^b \|\mathbf{f}(t)\|dt \geq \|\int_a^b \mathbf{f}(t)dt\|$

A.8 Funções de Múltiplas Variáveis

A.8.1 Exemplos

Seja $D \subseteq \mathbb{R}^2$ um subconjunto do plano xy e seja $f(x, y)$ uma função que associa a cada ponto (x, y) de D um escalar. Tal função é dita *função de duas variáveis*. O conjunto D é dito *domínio* de f. O conjuunto $I = \{f(x, y) : (x, y) \in D\}$ é dito *varredura* de f.

Dois exemplos:

- Para $D = \mathbb{R}^2$, a função $f(x, y) = xy$ associa um valor real a cada elemento de D.
- Seja $D = \{(x, y) : x^2 + y^2 < 1\}$ o disco aberto de raio unitário no plano xy. A função $f(x, y) = \frac{1}{\sqrt{1-(x^2+y^2)}}$ está definida dentro do disco aberto D.

Seja $D \subseteq \mathbb{R}^3$ um subconjunto do espaço xyz. Uma função $f(x, y, z)$ que associa um valor real a cada ponto (x, y, z) de D é dita *função de três variáveis*. Da mesma forma que no caso anterior, D é o domínio de f e $\{f(x, y, z) : (x, y, z) \in D\}$ é a *varredura*.

Como exemplo, Seja $D = \{(x, y, z) : x^2 + y^2 + z^2 < 1\}$ uma bola tridimensional aberta de raio unitário. A cada ponto $(x, y, z) \in D$ associamos o número $f(x, y, z) = \sqrt{1 - (x^2 + y^2 + z^2)}$.

A.8.2 Superfícies

As superfícies no espaço xyz definidas por equações da forma:

$$ax^2 + by^2 + cz^2 + dxy + exz + fyz + hx + iy + jz + k = 0 \qquad \text{(A.11)}$$

são conhecidas por superfícies quadráticas. Exemplos de superfícies desta família são: i) elipsóide; ii) hiperbolóide de uma folha; iii) hiperbolóide de duas folhas; iv) cónica quadrática.

A.8.3 Elipsóide

O elipsóide com centro na origem e simétrico com respeito aos três eixos consiste das soluções da equação:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

O elipsóide $\varepsilon = \{(x, y, z) : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\}$ intercepta os eixos coordenados em 6 pontos: $(\pm a, 0, 0)$, $(0, \pm b, 0)$, e $(0, 0, \pm c)$. Esses pontos são conhecidos por vértices.

A superfície definida por ε é limitada, satisfazendo: $|x| \leq a$, $|y| \leq b$, e $|z| \leq c$.

Todos os traços do elipsóide definem elipses. Por exemplo, o traço no plano xy é obtido fazendo $z = 0$ que consiste da equação:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

a qual define uma elipse no plano xy.

Todas as seções paralelas aos planos coordenados são também elipses. Por exemplo, fazendo $y = y_0$, teremos:

$$\frac{x^2}{a^2} + \frac{z^2}{c^2} = 1 - \frac{y_0^2}{b^2}$$

Esta elipse é a interseção do elipsóide com o plano definido por $y = y_0$.

Os números a, b e c são os semi-eixos do elipsóide. Quando todos os semi-eixos são iguais, a superfície ε define uma esfera.

A.8.4 Derivadas Parciais

Seja f uma função de x, y e z, por exemplo, considere $f(x, y, z) = 3x^2y - 5x\cos(\pi y) + \sin(z)x^2$. A *derivada parcial de f com respeito a x* é a função $\frac{\partial f}{\partial x}$ obtida diferenciando f com respeito a x, enquanto que y e z são tomadas como constantes. Para o exemplo dado,

$$\frac{\partial f}{\partial x} = 6xy - 5\cos(\pi y) + 2\sin(z)x$$

A *derivada parcial de f com respeito a y* é a função $\frac{\partial f}{\partial y}$ obtida diferenciando f com respeito a y, enquanto que x e z são tomadas como constantes. Para o exemplo dado,

$$\frac{\partial f}{\partial y} = 3x^2 + 5\pi x\sin(\pi y)$$

Da mesma forma para z, obtemos

$$\frac{\partial f}{\partial z} = x^2\cos(z)$$

Formalmente, as derivadas parciais são definidas em termos dos limites:

$$\begin{aligned}\frac{\partial f}{\partial x} &= \lim_{h\to 0}\frac{f(x+h,y,z)-f(x,y,z)}{h}\\ \frac{\partial f}{\partial y} &= \lim_{h\to 0}\frac{f(x,y+y,z)-f(x,y,z)}{h}\\ \frac{\partial f}{\partial z} &= \lim_{h\to 0}\frac{f(x,y,z+h)-f(x,y,z)}{h}\end{aligned}$$

Definição A.12 *Seja $f:\mathbb{R}^n\to\mathbb{R}$ uma função multivariada. Se as primeiras derivadas parciais de f são contínuas no ponto $\mathbf{x}$, então f é diferenciável em x e*

$$\nabla f(\mathbf{x}) = \frac{\partial f}{\partial x}(\mathbf{x})\mathbf{i} + \frac{\partial f}{\partial y}(\mathbf{x})\mathbf{j} + \frac{\partial f}{\partial z}(\mathbf{x})\mathbf{k}$$

A.8.5 Diferenciação e Gradiente

No caso de funções univariadas, o quociente

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

era dito derivada de f no ponto x quando o quociente tinha limite para $h\to 0$. No caso multivariado, podemos tomar a diferença $f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})$, mas o quociente

$$\frac{f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})}{\mathbf{h}}$$

não faz sentido já que h é um vetor.

Seja $o(\mathbf{h})$ um número que é proporcional ao comprimento de $\mathbf{h}$, i.e. $\|\mathbf{h}\|$. Dizemos que $f:\mathbb{R}^n\to\mathbb{R}$ é diferenciável no ponto $\mathbf{x}$ se e somente se existe um vetor $\mathbf{y}$ tal que

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \mathbf{y}\cdot\mathbf{h} + o(\mathbf{h})$$

Tal vetor $\mathbf{y}$, quando definido, é único. Este vetor é dito *gradiente de f no ponto* $\mathbf{x}$ e denotado por $\nabla f(\mathbf{x})$.

A.9 Conversão Entre Bases

A.9.1 Conversão de Números Inteiros

Sendo N um número inteiro expresso em uma base S, representado por N_S, deseja-se converter esse mesmo número à base R ($R\neq S$), ou seja transformar N_S em M_R mantendo-se o valor do dado. Para fazermos tal conversão

Tabela A.1: Exemplo de conversão de base: $(283)_{10} = (100011011)_2$

n	0	1	2	3	4	5	6	7	8	9
N_n	283	141	70	35	17	8	4	2	1	0
A_n	—	1	1	0	1	1	0	0	0	1

podemos lançar mão do método da sequência de divisões. Inicialmente dividiremos N por R, que gerará um resto A_1 e um quociente N_1, que depois de dividido por R gerará um resto A_2 e um quociente N_2. Repetiremos esse processo até chegarmos a um $N_{n-1} < R$, que produzirá um $A_n = N_{n-1}$. Uma vez feitas todas as divisões necessárias podemos construir o M, $M = A_n A_{n-1} \ldots A_1$. Esse resultado pode ser melhor entendido se repararmos que:

$$\begin{aligned} N &= R \times N_1 + A_1 \\ &= R \times (R \times N_2 + A_2) + A_1 \\ &= R \times (R \times (R \times N_3 + A_3) + A_2) + A_1 \\ &\quad\quad\quad\quad \vdots \\ &= R \times (R \times (R \times (\ldots (R \times N_{n-1} + A_{n-1}) \ldots) + A_2) + A_1 \\ &= R \times (R \times (R \times (\ldots (R \times A_n + A_{n-1}) \ldots) + A_2) + A_1 \end{aligned}$$

Ou seja:

$$N = A_n \times R^{n-1} + A_{n-1} \times R^{n-2} + \ldots + A_2 \times R + A_1$$

Comprovando que $N_S = M_R$.

Exemplo:

Deseja-se converter o número $(283)_{10}$ para a base binária.

De acordo com o método das divisões sequenciais, teremos a Tabela A.9.1

Lembrando que N_n é o quociente e A_n é o resto da divisão de N_{n-1} por R, no caso 2.

Uma vez calculados os valores de A_n podemos dizer que $(283)_{10} = (100011011)_2$, de acordo com $M = A_n A_{n-1} \ldots A_1$ e $283 = 1 \times 2^8 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^1 + 1 \times 2^0$.

A.9.2 Conversão de Números Puramente Fracionários

Como vimos, a conversão de um número inteiro N pode ser representada por: $N = A_n \times R^{n-1} + A_{n-1} \times R^{n-2} + \ldots + A_2 \times R + A_1$. Por extensão,

Tabela A.2: Exemplo de conversão de base: $(0.283)_{10} = (0.01001)_2$

i	$NF \times R$	$=$	$R \times NF$	A_i
-1	0.283×2	$=$	0.566	0
-2	0.566×2	$=$	1.132	1
-3	0.132×2	$=$	0.264	0
-4	0.264×2	$=$	0.528	0
-5	0.528×2	$=$	1.056	1

podemos dizer que a parte fracionária de um número (aqui chamada de NF) tem sua conversão para a base R dada por:

$$NF = A_{-1} \times R^{-1} + A_{-2} \times R^{-2} + A_{-3} \times R^{-3} + \ldots$$

Podemos definir os valores de A_{-1},A_{-2}, etc por uma sequência de multiplicações, ou seja ao pegarmos o NF e multiplicarmos por R, teremos um novo número, o qual sua parte inteira será igual ao fator A_{-1}. Se subtrairmos o valor de A_{-1} da multiplicação, teremos um novo número fracionário sem parte inteira, e podemos aplicar novamente o método. O número de iterações determinará o número de casas após a vírgula (precisão) da conversão.

$$\begin{aligned} NF &= A_{-1} \times R^{-1} + A_{-2} \times R^{-2} + A_{-3} \times R^{-3} + \ldots \\ NF \times R &= A_{-1} + A_{-2} \times R^{-1} + A_{-3} \times R^{-2} + \ldots \\ R \times (NF \times R - A_{-1}) &= A_{-2} + A_{-3} \times R^{-1} + \ldots \end{aligned}$$

A parte inteira de $NF \times R$ será o valor de A_{-1}. A parte inteira de $R \times (NF \times R - A_{-1})$ será o valor de A_{-2}. E assim sucessivamente até o número desejado de casas após a vírgula.

Exemplo:

Deseja-se converter o número $(0.283)_{10}$ para a base binária, com uma precisão de 5 casas após a vírgula. De acordo com o método das multiplicações sequenciais, teremos a Tabela A.9.2.

Uma vez montada a tabela é fácil visualizar os valores de A_i e por consequência chegar a conversão desejada: $(0.283)_{10} = (0.01001)_2$

A.9.3 Conversão de Números com Parte Fracionária e Parte Inteira

Caso um número contenha tanto parte inteira quanto fracionária, devemos proceder a conversão separadamente e juntar as duas soluções no final,

pois:

$$\begin{aligned} N_T &= N + NF \\ N &= A_n \times R^{n-1} + A_{n-1} \times R^{n-2} + \ldots + A_2 \times R + A_1 \\ NF &= A_{-1} \times R^{-1} + A_{-2} \times R^{-2} + A_{-3} \times R^{-3} + \ldots \end{aligned}$$

Portanto:

$$\begin{aligned} N_T = \quad & A_n \times R^{n-1} + A_{n-1} \times R^{n-2} + \ldots + A_2 \times R + A_1 \\ & + A_{-1} \times R^{-1} + A_{-2} \times R^{-2} + A_{-3} \times R^{-3} + \ldots \end{aligned}$$

Exemplo:

Deseja-se converter o número $(283.283)_{10}$ para a base binária, com uma precisão de 5 casas após a vírgula. Utilizando os resultados já encontrados nos exemplos anteriores, chegamos ao resultado: $(283.283)_{10} = (100011011.01001)_2$

A.9.4 Exercícios

i. Um dado número tem sua representação dada por $(275.65625)_{10}$, pede-se sua representação na base binária com precisão de 5 casas decimais.

ii. Desafio: Usando o MatLab (ou similar), implementar um algoritmo que transforme um certo número decimal em um binário, e vice versa.

A.10 Referências

Os conceitos apresentados neste apêndice são resumos editados e traduzidos do livro texto de Salas, Hille e Anderson [6]. O leitor interessando em aprofundar os tópicos acima pode consultar diretamente o referido texto.

Apêndice B

Exemplos de Código Matlab

B.1 Capítulo 3

B.1.1 Figura 3.2

```
x =[];
y1 = [];
y2 = [];
y3 = [];
y4 = [];
y5 = [];
j = 0;
for k=-3:0.1:3
   j = j+1;
   x(j) = k;
   y1(j) = exp(k);
   y2(j) = exp(k) - exp(-k);
   y5(j) = 0;
end

clf;
subplot(2,2,1);
plot(x,y1);
hold;
plot(x,y5);

subplot(2,2,2);
plot(x,y2);
hold;
plot(x,y5);

j=0;
for k=-2:0.1:2
   j = j+1;
```

```
   x3(j) = k;
   y3(j) = exp(k) - exp(-k) -3*k;
end
subplot(2,2,3);
plot(x3,y3);
hold;
plot(x,y5);

x4 = [];
y4b = [];
j=0;
for k=-5:0.1:5
   j = j+1;
   x4(j) = k;
   y4(j) = cos(k) - 0.5;
   y4b(j) = 0;
end
subplot(2,2,4);
plot(x4,y4);
hold;
plot(x4,y4b);
```

B.1.2 Método de Ponto Fixo para Função $f(x) = x^2/10 - x + 1$

```
%--------------------------
% Iterative Process

x = [];
l = [];
x(1) = 1.5;
l(1) = 1;
for k=2:100
   x(k) = (x(k-1)^2)/10 + 1;
   l(k) = k;
end

plot(l,x);
```

B.1.3 Algoritmo da Bisecção

```
function [x] = bissec(funcao,a,b);

%    BISSEC calcula a raiz de funçoes nao-lineares dentro de um
% intervalo estabelecido. Ela recebe os seguintes parametros:
%
%     -> BISSEC(funcao,a,b)
```

```
%
%       - funcao = funcao definida por apenas uma variavel qualquer. Deve estar entre
%       aspas simples.
%       - a = extremo esquerdo do intervalo de avaliaçao
%       - b = extremo direito do intervalo de avaliaçao
%
%     Exemplo:
%
%         >> BISSEC('2*x^2+3*x-6',1,3)
%
%         ans %
%             1.1375


if (nargin ~= 3)
    error('A funçao BISSEC precisa de tres parametros. Mais informaçoes: >> help bissec');
end

lim=1000;
e1=10^(-10);
e2=10^(-10);
x0=a;
x1=b;
i=0;
f0=subs(funcao,x0);
f1=subs(funcao,x1);

if (f0*f1 > 0)
    error('Nao existe raiz entre a e b');
end

while (i < lim)

    if (abs(f0) <= e2)
        x=x0;
        break
    end
    if (abs(f1) <= e2)
        x=x1;
        break
    end
    if ((abs(x0-x1)) < (e1*(abs(x1))))
        x=x0;
        break
    end

    x2=(x0+x1)/2;
    f2=subs(funcao,x2);
```

```
    if ((f2*f0) < 0)
        x1=x2;
        f1=f2;
    else
        x0=x2;
        f0=f2;
    end

    i=i+1;
end

if (i > lim)
    error('Nao atingiu exatidao');
end
```

Apêndice C

Exercícios Resolvidos

Capítulo 2

Seção 2.1

i.

$$\begin{aligned} G &= \frac{F}{E} \\ &= \frac{(Y+Y+Y+Y+Y)-H}{(X+X+X)-H} \\ &= \frac{5\left(\frac{3}{5}-H\right)-H}{3\left(\frac{2}{3}-H\right)-H} \\ &= \frac{3-6H}{2-4H} \\ &= \frac{0}{0} \end{aligned}$$

Indeterminação.

ii. (a) HP25, $F(10, 9, -98, 100)$

(a) 358200000001

(b) $0.100000000 \cdot 10^{-98}$

(c) $0.999999999 \cdot 10^{100}$

(d) $\{x : |x| > 0.999999999 \cdot 10^{100}\}$

(e) $\{x : 0 < |x| < 0.100000000 \cdot 10^{-98}\}$

(b) IBM 360/370, $F(16, 6, -64, 63)$

(a) 4.026531841

(b) $0.100000 \cdot 16^{-64}$

(c) $0.FFFFFF \cdot 16^{63}$

(d) $\{x : |x| > 0.FFFFFF \cdot 16^{63}\}$

(e) $\{x : 0 < |x| < 0.100000 \cdot 16^{-64}\}$

(c) B6700, $F(8, 13, -51, 77)$

(a) 124107374985217

(b) $0.1000000000000 \cdot 8^{-51}$

(c) $0.7777777777777 \cdot 10^{77}$

(d) $\{x : |x| > 0.7777777777777 \cdot 10^{77}\}$

(e) $\{x : 0 < |x| < 0.1000000000000 \cdot 8^{-51}\}$

iii. 0.110 para ambos os casos

Seção 2.4.1

i. $\nabla(a) = \frac{5}{8} \quad \Delta(a) = \frac{3}{4} \quad o(a) = \frac{3}{4} \qquad \nabla(b) = \frac{5}{8} \quad \Delta(b) = \frac{3}{4} \quad o(b) = \frac{5}{8}$

ii. $E_A(\nabla(a)) = \frac{5}{56} \qquad E_R(\nabla(a)) = \frac{1}{8}$

$E_A(\Delta(a)) = \frac{1}{28} \qquad E_R(\Delta(a)) = \frac{1}{20}$

$E_A(o(a)) = \frac{5}{56} \qquad E_R(o(a)) = \frac{1}{8}$

$E_A(\nabla(b)) = \frac{1}{24} \qquad E_R(\nabla(b)) = \frac{1}{16}$

$E_A(\Delta(b)) = \frac{1}{12} \qquad E_R(\Delta(b)) = \frac{1}{8}$

$E_A(o(b)) = \frac{1}{24} \qquad E_R(o(b)) = \frac{1}{16}$

iii. $DIGSE(x) = 4$
$DIGSE(y) = 4$

iv. Sim

Seção 2.8

Exercício 2.1 Não. Aumento de precisão não implica aumento de exatidão.

Exercício 2.2 Não. Há muitos algoritmos mais exatos porém com custos temporais maiores para chegar a solução.

Exercício 2.3 Erro Inerente: Erro intrínseco ao modelo matemático adotado. Para corrigi-lo, faz-se necessária uma revisão do modelo identificando as origens de tal erro.
Erro de Discretização: Erro que aparece quando substitui-se um processo infinito por um processo finito. Para diminui-lo, deve ser aumentado o número de iterações ou termos relevantes.
Erro de Arredondamento: Surge ao trabalharmos com máquinas digitais e sistemas de ponto flutuante, onde faz-se necessário o arredondamento. Pode ser minimizado utilizando o arredondamento para o número de máquina mais próximo (ox).

Exercício 2.6 17

Exercício 2.7 28.8125; 0.421875; 1000011; 1011101.001

Exercício 2.10

$$\begin{aligned}
y_n &= \int_0^2 n^{n\log_n x} e^{x-1}dx = \int_0^2 x^n e^{x-1}dx \\
y_1 &= \int_0^2 xe^{x-1}dx = \frac{1+e^2}{e} \\
y_n &= \int_0^2 x^n e^{x-1}dx = 2^n e - ny_{n-1} \\
y_1 &= \frac{1+e^2}{e} \\
y_2 &= 2^2 e - 2y_1 = \frac{2(e^2-1)}{e} \\
y_3 &= 2^3 e - 3y_2 = \frac{2(e^2+3)}{e} \\
y_4 &= 2^4 e - 4y_3 = \frac{8(e^2-3)}{e}
\end{aligned}$$

Exercício 2.11 (a) $\{\, 0 < |\, x| < 0.250 \,\}$
(b) $\{\, |\, x| > 3.5 \,\}$
(c) 0.250

(d) 0.3125

(e) 0.375

Exercício 2.12 b_1 precisa ser maior que zero pois se fosse zero teríamos duas representações diferentes em ponto flutuante para o mesmo número, o que é não desejável.

Exercício 2.13 i. Sim

ii. Não

iii. Sim

iv. Sim

Exercício 2.14 Item a)

$$\begin{aligned}
\tilde{f}(x+\Delta x) &= \sqrt{x}\sin(x) + \left[\frac{1}{2\sqrt{x}}\sin(x) + \sqrt{x}\cos(x)\right]\Delta x + \\
&\quad + \left[\frac{-1}{4(x)^{\frac{3}{2}}}\sin(x) + \frac{\cos(x)}{2\sqrt{x}} + \frac{1}{2\sqrt{x}}\cos(x) - \sqrt{x}\sin(x)\right]\frac{\Delta x^2}{2} \\
\tilde{f}(x) &= \sqrt{x_0}\sin(x_0) + \left[\frac{1}{2\sqrt{x_0}}\sin(x_0) + \sqrt{x_0}\cos(x_0)\right](x-x_0) + \\
&\quad + \left[\frac{-1}{4(x_0)^{\frac{3}{2}}}\sin(x_0) + \frac{\cos(x_0)}{2\sqrt{x_0}} + \frac{1}{2\sqrt{x_0}}\cos(x_0) - \sqrt{x_0}\sin(x_0)\right]\frac{(x-x_0)^2}{2} \\
&= \sqrt{\frac{\pi}{2}}\sin\left(\frac{\pi}{2}\right) + \left[\frac{1}{2\sqrt{\frac{\pi}{2}}}\sin\left(\frac{\pi}{2}\right) + \sqrt{\frac{\pi}{2}}\cos\left(\frac{\pi}{2}\right)\right]\left(x-\frac{\pi}{2}\right) + \\
&\quad \left[\frac{-1}{4\left(\frac{\pi}{2}\right)^{\frac{3}{2}}}\sin\left(\frac{\pi}{2}\right) + \frac{\cos\left(\frac{\pi}{2}\right)}{2\sqrt{\frac{\pi}{2}}} + \frac{1}{2\sqrt{\frac{\pi}{2}}}\cos\left(\frac{\pi}{2}\right) - \sqrt{\frac{\pi}{2}}\sin\left(\frac{\pi}{2}\right)\right]\frac{\left(x-\frac{\pi}{2}\right)^2}{2} \\
&= 1.2533 + 0.3985\left(x-\frac{\pi}{2}\right) - 1.3803\frac{(x-\frac{\pi}{2})^2}{2}
\end{aligned}$$

Item b) $\tilde{f}\left(\frac{\pi}{3}\right) = 0.8552$. **Item c)** $f\left(\frac{\pi}{3}\right) = 0.8862,\ E_R = 3.50$.

Exercício 2.15 Conversão de números na base 4 para base 5 em ponto flutuantes. Tem-se os algorismos na base a ser transformada:

$$N_4 = 0.132$$

Para facilitar o processo converte-se esse número à uma base familiar(base 10, por exemplo):

$$N_4 \to N_{10} = 0\times 4^0 + 1\times 4^{-1} + 3\times 4^{-2} + 2\times 4^{-3} = 0 + \frac{1}{4} + \frac{3}{16} + \frac{2}{64} =$$

$$\frac{8+6+1}{32} = \frac{15}{32} = 0,46875$$

A partir do número obtido, aplica-se o método de conversão já conhecido(base 10 para outra base qualquer):

i. O número escrito na base 10 é multiplicada pelo número de algorismos possíveis na nossa base(5, para base 5, 2 para base 2);

ii. Do número resultante da multiplicação subtrai-se o maior número inteiro possível que ainda o mantenha positivo;

iii. Esse número inteiro é o primeiro elemento da mantissa do algorismo já convertido e a mantissa resultante da subtração passa novamente pelos processos 1,2,3.

$$\left.\begin{array}{l} 0,46873 * 5 \rightarrow 2,34375 \Rightarrow 2 \\ 0,34375 * 5 \rightarrow 1,71875 \Rightarrow 1 \\ 0,71875 * 5 \rightarrow 3,59375 \Rightarrow 3 \\ 0,59375 * 5 \rightarrow 2,96875 \Rightarrow 2 \\ 0,96875 * 5 \rightarrow 3,84375 \Rightarrow 4 \end{array}\right\}$$

Logo, o número $N_4 = 0,132$, $N_{10} = 0,46875$ em base 5 é $0,21324^*$...

$$\frac{1}{5} + \frac{1}{25} + \frac{1}{125} + \frac{1}{625} + \frac{1}{3125}$$

*O número obtido através desse método possui "infinitos"algorismo, visto que poemos multiplicar "infinitamente"a mantissa por 5^4, logo o valor resultante, no seu reconvertidos será uma aproximação do valor real. Entretanto, existem casos em que a mantissa assume o valor zero, logo nesse caso o número tem dígitos finitos.

Capítulo 3

Seção 3.9

Exercício 3.1 Discordo. Há casos onde o Método de Newton não converge para a solução entrando em laços infinitos ou divergindo.

Exercício 3.2

$$\begin{aligned} x^{k+1} &= x^k - \frac{f(x^k)}{f'(x^k)} = x^k - \frac{\sqrt{x^k}}{\frac{1}{2\sqrt{x^k}}} = x^k - 2x^k = -x^k \qquad se\ \ x > 0 \\ x^{k+1} &= x^k - \frac{f(x^k)}{f'(x^k)} = x^k + \frac{\sqrt{-x^k}}{\frac{1}{2\sqrt{-x^k}}} = x^k - 2x^k = -x^k \qquad se\ \ x < 0 \end{aligned}$$

a) Sim. $x = \{0\}$

b) Não

c) Sim. $x = \Re - \{0\}$

Exercício 3.3 Discordo.

Exercício 3.4 $x^* = 0.56116456934659$

Exercício 3.5 $x = 1.8$

Exercício 3.6

$$\begin{aligned} (x - x^2)^2 - 3x^2 - 5x - 5 &= 0 \\ x^4 - 2x^3 - 2x^2 - 5x - 5 &= 0 \end{aligned}$$

Aplicando-se o Método da Bissecção na equação acima com $(a, b) = (-1.5, 0)$ tem-se como solução $z^* = (-0.9006426, -1.7117997)$.

Exercício 3.7

$$\begin{aligned} g(x) &= x + c(x)f(x) \\ g(x) &= x + (xe^{-x})\left(-\frac{1}{e^{-x}(1+x)}\right) \\ g(x) &= \frac{x^2}{x+1} \end{aligned}$$

Para o ponto fixo convergir para uma solução com $x_0 \in I = [0, 1)$, a função deve obedecer a três condições:

i. Deve ser contínua em I.

Visivelmente satisfeita.

ii. $g(I) \in I$

Como $g(0) = 0$ e $g(1) = 0.5$ e todos os outros pontos de $g(I)$ estão dentro de I, a condição 2 é satisfeita.

iii. $|g'(I)| < 1$

$$g'(x) = \frac{x^2 + 2x}{(x+1)^2}$$

Como $|g'(0)| = 0$ e $|g'(1)| = 0.75$ e todos os outros pontos de $g'(I)$ estão dentro de I, a condição 3 é satisfeita.

Exercício 3.8 Pode-se calcular a raiz de um número x_s pelo método de Newton.

$$\begin{aligned} y &= \sqrt{x_s} \\ y^2 &= x_s \\ y^2 - x_s &= 0 \end{aligned}$$

Aplicando o Método de Newton na equação $f(y) = y^2 - x_s$ acima visando encontrar uma solução positiva y_n positiva:

$$\begin{aligned} y_{n+1} &= y_n - \frac{f(y_n)}{f'(y_n)} \\ y_{n+1} &= y_n - \frac{y_n^2 - x_s}{2y_n} \\ y_{n+1} &= \frac{1}{2}\left(y_n + \frac{x_s}{y_n}\right) \end{aligned}$$

Se iterar-se esta função a partir de um chute inicial y_0 próximo à solução, o método irá convergir para a solução ideal.

Exercício 3.9 Implemente em uma plataforma computacional.

Exercício 3.10 $\nabla F(x,y) = \begin{bmatrix} \partial f_1/\partial x & \partial f_1/\partial y \\ \partial f_2/\partial x & \partial f_2/\partial y \end{bmatrix}$

$$\nabla F = \begin{bmatrix} e^{-x^2-y^2} - 2xe^{-x^2-y^2}(x-y+\frac{1}{4}) & -e^{-x^2-y^2} - 2ye^{-x^2-y^2}(x-y+\frac{1}{4}) \\ 4x & -\sin y \end{bmatrix}$$

$$\nabla F(1,0) = \begin{bmatrix} -0.55181916 & -0.36787944 \\ 4 & 0 \end{bmatrix}$$

$$F(1,0) = \begin{bmatrix} 0.45984930 \\ 3 \end{bmatrix}$$

Para calcular o primeiro iterando utiliza-se a fórmula:

$$\begin{aligned} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} &= \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \nabla F^{-1}(x_0, y_0) \cdot F(x_0, y_0) \\ &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0.25 \\ -2.71828183711 & -0.375 \end{bmatrix} \cdot \begin{bmatrix} 0.45984930 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 0.25 \\ 2.375 \end{bmatrix} \end{aligned}$$

Exercício 3.11

i) Falso

ii) Falso

iii) Falso

iv) Falso

v) Falso

Exercício 3.12 Substituindo as expressões:

$$\begin{aligned} f(x) &= x^3 + 2.21x + 2(5.3e^x - x) \\ &= x^3 + 0.21x + 10.6e^x \end{aligned}$$

Basta aplicar algum método numérico para encontrar um x^* tal que $f(x^*) = 0$.

Exercício 3.13

$$P(\lambda) = -\lambda^5 + 2\lambda^4 + 72\lambda^3 + 34\lambda^2 - 647\lambda - 918$$

$$autovalores(A) = \{-2.2307,\ -2,\ -6.4588,\ 3.4468,\ 9.2427\}$$

Exercício 3.14 Desenvolver em uma plataforma computacional.

Exercício 3.15 O problema consiste em:

$$\begin{aligned} \min \;\; P(x,y) &= \min \;\; \pi x + 2x + 2y \\ A(x,y) &= 2xy = 3.8 \end{aligned}$$

Fazendo as devidas substituições:

$$\begin{aligned} f(x) &= \pi x + 2x + \frac{3.8}{x} \\ f'(x) &= \pi + 2 - \frac{3.8}{x^2} \\ f''(x) &= \frac{7.6}{x^3} \end{aligned}$$

Note que para $x > 0$, $f''(x)$ está crescendo, então $f(x)$ é convexa. Portanto, o ponto de mínimo pode ser encontrado fazendo $f'(x) = 0$.

$$\begin{aligned} f'(x) &= 0 \\ \pi + 2 - \frac{3.8}{x^2} &= 0 \\ x &= \sqrt{\frac{3.8}{\pi + 2}} \end{aligned}$$

Tem-se que $x = 0.859694$ e $y = 2.210088$.

Exercício 3.17 O processo iterativo advém da aplicação do método de Newton para computar $f(x) = \sqrt{x}$.

a) Primeiramente, observe que $g(y)$ é continua em I e portanto nos resta mostrar que $g(y) \in I$ para todo $y \in I$.

$$\begin{aligned} g(y) \in I &\Leftrightarrow 1 < g(y) < x \\ &\Leftrightarrow 1 < \frac{y^2 + x}{2y} < x \end{aligned}$$

$1 < (y^2 + x)/2y \Leftrightarrow 2y < y^2 + x \Leftrightarrow y^2 - 2y + x > 0$. Analisando as raízes da função quadrática, temos que $\Delta = b^2 - 4ac = 4 - 4x < 0 \Leftrightarrow x > 1$. Portanto a quadrática tem somente raízes complexas e sendo convexa nunca assumirá valor negativo, verificando-se assim que $1 < (y^2 + x)/2y$ para todo $y \in I$.

Por outro lado, $(y^2 + x)/2y < x \Leftrightarrow y^2 + x < 2yx \Leftrightarrow y^2 - 2xy + x < 0$. Analisando as raízes da função quadrática, temos que $\Delta = b^2 - 4ac = (-2x)^2 - 4x = 4x^2 - 4x = 4x(x-1)$. $\Delta > 0 \Leftrightarrow$

$4x^2 - 4x > 0 \Leftrightarrow 4x^2 > 4x \Leftrightarrow x^2 > x \Leftrightarrow |x| > 1$, que é verificado já que $x > 1$. Considere as duas raízes da quadrática dadas por:

$$\begin{aligned} y' &= \frac{-b + \sqrt{\Delta}}{2a} = \frac{2x + \sqrt{4(x^2 - x)}}{2} \\ &= x + \sqrt{x^2 - x} \\ &> x \quad (\text{já que } x > 1) \\ y'' &= \frac{-b - \sqrt{\Delta}}{2a} = \frac{2x - \sqrt{4(x^2 - x)}}{2} \\ &= x - \sqrt{x^2 - x} \end{aligned}$$

Note que:

$$\begin{aligned} y'' > 1 &\Leftrightarrow x - \sqrt{x^2 - x} > 1 \\ &\Leftrightarrow (x - 1) > \sqrt{x^2 - x} \\ &\Rightarrow (x - 1)^2 > \sqrt{x^2 - x}^2 \\ &\Leftrightarrow x^2 - 2x + 1 > x^2 - x \\ &\Leftrightarrow -2x + 1 > -x \\ &\Leftrightarrow x < 1 \end{aligned}$$

Assim, a raiz $y'' \leq 1$ e a raiz $y' > x$. Já que a quadrática $y^2 - 2xy + x < 0$ é convexa, na região definida pelo intervalo $I = (1, x)$ esta assume valor estritamente negativo e, portanto, $(y^2 + x)/2y < x$. O desenvolvimento acima mostra que $g(y)$ é contínua e $g(I) \subseteq I$, garantindo a existência de um ponto fixo em I.

Capítulo 4

Seção 4.19

Exercício 4.1

a) Aplicando na *equação Diofantina*, tem-se:

$$(a_0+a_1s)(d_0+d_1s+d_2s^2)+(b_0+b_1s)(n_0+n_1s+n_2s^2) = \delta_0+\delta_1s+\delta_2s^2+\delta_3s^3$$

Por igualdade de polinômios:

$$\begin{cases} a_0d_0 + b_0n_0 = \delta_0 \\ a_1d_0 + b_0d_1 + b_1n_0 + b_0n_1 = \delta_1 \\ a_0d_2 + a_1d_1 + b_0n_2 + b_1n_1 = \delta_2 \\ a_1d_2 + b_1n_2 = \delta_3 \end{cases}$$

Escrevendo em modo matricial:

$$\begin{bmatrix} d_0 & 0 & n_0 & 0 \\ d_1 & d_0 & n_1 & n_0 \\ d_2 & d_1 & n_2 & n_1 \\ 0 & d_2 & 0 & n_2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

b) Pelos dados fornecidos:

$$\begin{cases} n_0 = 1 & n_1 = 1 & n_2 = 0 \\ d_0 = 1 & d_1 = 2.5 & d_2 = 1 \\ \delta_0 = 40 & \delta_1 = 34 & \delta_2 = 10 & \delta_3 = 1 \end{cases}$$

Substituindo na matriz:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 2.5 & 1 & 1 & 1 \\ 1 & 2.5 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 40 \\ 34 \\ 10 \\ 1 \end{bmatrix}$$

Após aplicar o Método de Gauss com pivoteamento:

$$\begin{aligned} A(s) &= -29 + s \\ B(s) &= 69 + 36.5s \end{aligned}$$

c)

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 40 \\ 33 \\ 7.5 \end{bmatrix}$$

Exercício 4.2

a) O critério das linhas diz que dado um sistema de equações $Ax = b$, se

$$\sum_{j=1\,;\,j\neq i}^{n} |\,a_{ij}| < |\,a_{ii}| \quad ,\ i = 1,\ 2, ..., n$$

então o Método de Gauss-Seidel gera uma seqüência convergente.

Aplicando-o para as linhas da matriz:

1^a linha) $|0.7| + |-1| + |1| < |3|$ Verdadeiro
2^a linha) $|1| + |-1| + |0| < |2|$ Falso
3^a linha) $|1| + |-1| + |0.7| < |3|$ Verdadeiro
4^a linha) $|1| + |0| + |1| < |2|$ Falso

Não se pode afirmar nada sobre os Métodos de Jacobi e Gauss-Seidel a partir do critério das linhas pois a condição suficiente não foi corroborada.

b) Aplicando o critério de Sassenfeld obteve-se os seguintes β:

$$\beta_1 = 0.9 \qquad \beta_{21} = 0.45 \qquad \beta_3 = 0.45 \qquad \beta_4 = 0.675$$

Por este critério, pode-se concluir que há convergência garantida.

c)

$$\begin{bmatrix} t_{11} & t_{21} & v_{11} & v_{21} \end{bmatrix}^t = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & -\frac{1}{18} & -\frac{5}{36} \end{bmatrix}^t$$

Exercício 4.3

a) Sim. Vide Teorema 4.5.

b) Sim. A convergência é uma característica intrínseca ao sistema, independente do chute inicial. Se converge para um ponto inicial diferente da solução, convergirá para todos.

Exercício 4.4

$$\begin{aligned}
\frac{||x - x'||}{||x||} &\leq \kappa(A)\frac{||b - b'||}{||b||} \\
||x - y|| &\leq \kappa(A)||x||\frac{||b - (b - \Delta b)||}{||b||} \\
||x - y|| &\leq \kappa(A)||x||\frac{||\Delta b||}{||b||}
\end{aligned}$$

$$S = \left\{ y \in \Re^n : ||x - y|| \leq \kappa(A)||x||\frac{||\Delta b||}{||b||} \right\}$$

Exercício 4.5 O método de Jacobi pode portar-se como o método de Gauss-Seidel. Para tanto, o chute inicial x_0 for igual a solução x^*.

Exercício 4.6 A solução mais sensata é a fatoração LU da matriz A, o que facilita a resolução de diferentes problemas com matrizes b diferentes. Caso optasse por outros métodos, como Eliminação de Gauss, a matriz A teria que ser processada a cada modelo diferente.

Exercício 4.7 Falso. O método iterativo pode ou não convergir caso $\det(A) = 0$. Para tanto, devem ser analisados os outros critérios como critério das linhas, diagonal dominante ou critério de Sassenfeld.

Exercício 4.8 Desenvolver sobre uma plataforma computacional. A solução é:

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} -0.6126 & 1.9404 & 1.5985 & 1.2270 \end{bmatrix}$$

Exercício 4.9 $||A||_\infty = 24 \qquad ||B||_1 = 40 \qquad ||C|| = 4$

Exercício 4.10 A matriz $B = -D^{-1}(A - D)$ define o processo iterativo de Jacobi com preparador clássico. Note que $\|B\|_1 = 1.1$ e $\|B\|_\infty = 1.1714$, não permitindo concluir que o processo converge. No entanto, $\|B\|_2 = \sqrt{\lambda_{max}(B^T B)} = 0.9530 < 1$, garantindo que o processo iterativo $x^{k+1} = Bx^k + D^{-1}b$ é convergente.

Exercício 4.11

i. $\kappa(A) = ||A||.||A^{-1}||$

ii. Não. O sistema terá obrigatoriamente infinitas soluções, não podendo ter apenas uma.

iii. Sim. Um sistema linear tem uma solução, infinitas soluções ou nenhuma solução.

iv. Correto.

Exercício 4.12

i. $rank(A) = rank([A|b]) < n$

ii. A matriz A deve ser definida positiva.

iii. $U = \begin{bmatrix} 1 & 1 & 3 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$

Para b_1: $x_1 = \begin{bmatrix} 7.25 & 2.75 & -3 \end{bmatrix}$

Para b_2: $x_2 = \begin{bmatrix} 16.25 & 6.75 & -7 \end{bmatrix}$

Exercício 4.13

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = CBA \begin{bmatrix} \Delta S_x \\ \Delta S_y \end{bmatrix}$$

onde: $C = \begin{bmatrix} \sin\theta_1 & \sin\theta_2 & -\cos\theta_3 \\ -\cos\theta_1 & \cos\theta_2 & -\cos\theta_3 \end{bmatrix}$

$$B = \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{bmatrix}$$

$$A = \begin{bmatrix} \sin\theta_1 & -\cos\theta_1 \\ \sin\theta_2 & \cos\theta_2 \\ -\cos\theta_3 & -\sin\theta_3 \end{bmatrix}$$

Resolvendo o sistema para as condições especificadas, chega-se à:

$$\Delta S = \begin{bmatrix} 0.0162 \\ 0.0275 \end{bmatrix}$$

Exercício 4.14

i. FALSO

ii. VERDADEIRO

iii. VERDADEIRO

iv. VERDADEIRO

v. FALSO

vi. FALSO

vii. FALSO

viii. VERDADEIRO

Exercício 4.15

$$Av_n = \lambda_n v_n$$
$$(A - \lambda_n I)v_n = 0$$

$$\begin{bmatrix} A-\lambda_1 I & 0 & \cdots & 0 \\ 0 & A-\lambda_2 I & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & A-\lambda_n I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Exercício 4.16 Falso. Testar os diferentes preparadores para o sistema:

$$\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

O sistema com preparador unitário não convirgirá enquanto o sistema com preparador clássico convirgirá.

Capítulo 5

Seção 5.10

Exercício 5.1

$$\begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 & 0 & 0 & 0 & 0 \\ 1 & t_2 & t_2^2 & t_2^3 & 0 & 0 & 0 & 0 \\ 1 & t_3 & t_3^2 & t_3^3 & 0 & 0 & 0 & 0 \\ 1 & t_4 & t_4^2 & t_4^3 & -1 & -t_4 & -t_4^2 & -t_4^3 \\ 0 & 1 & 2t_4 & 3t_4^2 & 0 & -1 & -2t_4 & -3t_4^2 \\ 0 & 0 & 0 & 0 & 1 & t_5 & t_5^2 & t_5^3 \\ 0 & 0 & 0 & 0 & 1 & t_6 & t_6^2 & t_6^3 \\ 0 & 0 & 0 & 0 & 1 & t_7 & t_7^2 & t_7^3 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ d_0 \\ d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 0 \\ 0 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix}$$

Exercício 5.2

$$\begin{bmatrix} x_{1,\,k+1} \\ x_{2,\,k+1} \end{bmatrix} = \begin{bmatrix} x_{1,\,k} \\ x_{2,\,k} \end{bmatrix} - \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{bmatrix}^{-1} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

onde:

$$\begin{aligned} \partial f_1/\partial x_1 &= \frac{2x_{1,\,k}}{(x_{1,\,k}^2 + 2x_{2,\,k}^2 + 1)\log(10)} \\ \partial f_1/\partial x_2 &= \frac{4x_{2,\,k}}{(x_{1,\,k}^2 + 2x_{2,\,k}^2 + 1)\log(10)} \\ \partial f_2/\partial x_1 &= -2x_{1,\,k} \\ \partial f_2/\partial x_2 &= 1 \\ f_1 &= \log(x_{1,\,k}^2 + 2x_{2,\,k}^2 + 1) - \frac{1}{2} \\ f_2 &= x_{2,\,k} - x_{1,\,k}^2 + 0.2 \end{aligned}$$

Exercício 5.3

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} - \begin{bmatrix} \partial f_1/\partial x & \partial f_1/\partial y & \partial f_1/\partial z \\ \partial f_2/\partial x & \partial f_2/\partial y & \partial f_2/\partial z \\ \partial f_3/\partial x & \partial f_3/\partial y & \partial f_3/\partial z \end{bmatrix}^{-1} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}$$

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} - \begin{bmatrix} 2x_k & 6z_k y_k^2 + 4y_k^3 & 2y_k^3 \\ 9x_k^2 - 12x_k z_k & -3y_k^2 & -6x_k^2 \\ 1 & -1 & -2 \end{bmatrix}^{-1} \begin{bmatrix} x_k^2 + 2z_k y_k^3 + y_k^4 \\ 3x_k^3 - 6x_k^2 z_k - y_k^3 \\ x_k - 2z_k - y_k \end{bmatrix}$$

Exercício 5.4 Faz sentido utilizar P quando $rank(A) = rank([A|b]) < n$, ou seja, tem-se um sistema subdimensionado com infinitas soluções. São aplicáveis em sistemas de otimização onde procura-se a solução de menor norma.

Exercício 5.5 Faz sentido utilizar $\hat{P}$ quando $rank(A) < rank([A|b])$, ou seja, não há solução para o sistema. Neste caso tenta-se minimizar o erro da solução encontrada. Tem aplicações em problemas similares como ajuste de curvas, ajuste de polinômios e identificação de sistemas discretos por predição.

Exercício 5.7

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ \frac{39}{2} & \frac{37}{2} & \cdots & \frac{3}{2} & \frac{1}{2} \end{bmatrix}_{2\times 20} \begin{bmatrix} v_1^x \\ v_2^x \\ \vdots \\ v_{19}^x \\ v_{20}^x \end{bmatrix}_{20\times 1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ \frac{39}{2} & \frac{37}{2} & \cdots & \frac{3}{2} & \frac{1}{2} \end{bmatrix}_{2\times 20} \begin{bmatrix} v_1^y \\ v_2^y \\ \vdots \\ v_{19}^y \\ v_{20}^y \end{bmatrix}_{20\times 1} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}$$

Exercício 5.8

a) 13.1816 e 0.1517

b) 13.1816

c) 13.1816

Exercício 5.9 É possível.

$$C_j = c_j \quad \Leftrightarrow \quad \ln C_j = \ln c_j \quad \Leftrightarrow \quad \ln(e^x a_j^y b_j^k g_j^z d_j^t) = \ln c_j$$

$$x + y \ln a_j + k \ln b_j + z \ln g_j + t \ln d_j = \ln c_j$$

O problema se constitui então em $\min ||Aw - b||^2$, aonde

$$A = \begin{bmatrix} 1 & \ln a_1 & \ln b_1 & \ln g_1 & \ln d_1 \\ 1 & \ln a_2 & \ln b_2 & \ln g_2 & \ln d_2 \\ & & \vdots & & \\ 1 & \ln a_m & \ln b_m & \ln g_m & \ln d_m \end{bmatrix} \qquad w = \begin{bmatrix} x \\ y \\ k \\ z \\ t \end{bmatrix} \qquad b = \begin{bmatrix} \ln c_1 \\ \ln c_2 \\ \vdots \\ \ln c_n \end{bmatrix}$$

Exercício 5.11

i. FALSO
ii. VERDADEIRO
iii. FALSO
iv. FALSO
v. VERDADEIRO
vi. VERDADEIRO
vii. FALSO
viii. FALSO
ix. VERDADEIRO

Exercício 5.12 Se $g(x) = -f''(x)$ e $h(x) = f'(x)$ então existe x tal que:

$$\begin{cases} g(x) < 0 & \Leftrightarrow \quad f''(x) > 0 \\ h(x) = 0 & \Leftrightarrow \quad f'(x) = 0 \end{cases}$$

Então x satisfaz a condição suficiente de segunda ordem para mínimo local. Então pode-se aplicar o pacote de software.

Capítulo 6

Seção 6.5

Exercício 6.1

a) Calculando a Cota de Laguerre-Thibault:

$$\begin{aligned} p(x) &= x^7 + 4x^6 - 7x^5 - 34x^4 - 24x^3 - x + 1 \\ S &= (+,+,-,-,-,-,+) \quad \Rightarrow \quad T = 2 \\ p(-x) &= -x^7 + 4x^6 + 7x^5 - 34x^4 + 24x^3 + x + 1 \\ S' &= (-,+,+,-,+,+,+) \quad \Rightarrow \quad T' = 3 \end{aligned}$$

Possui 2 ou nenhuma raízes positivas e, 3 ou 1 raízes negativas.

Por $a_5 = 0$, pode-se concluir pela Regra da Lacuna que para haver raízes complexas $a_4 \dot{a}_6 > 0$.

$$a_4 \cdot a_6 > 0 \quad \Rightarrow \quad (-24) \cdot (-1) > 0 \quad \Rightarrow \quad 24 > 0$$

Possui raízes complexas conjugadas.

VERDADEIRO

b) Pela Cota de Kojima, tem-se a seguinte série de fatores: $\{4, 2.646, 3.24, 2.213, 1, 1\}$. A soma dos dois maiores fatores é 7.24. Desta forma, pode-se afirmar que o módulo de toda e qualquer raiz de $p(x)$ é menor que 7.24.

VERDADEIRO

Exercício 6.2

$$\begin{aligned} p(x) &= 2x^7 + 4x^6 - 7x^5 + 12x^4 - x - 1 \\ S &= (+,+,-,+,-,-) \quad \Rightarrow \quad T = 3 \\ p(-x) &= -2x^7 + 4x^6 + 7x^5 + 12x^4 + x - 1 \\ S' &= (-,+,+,+,+,-) \quad \Rightarrow \quad T' = 2 \end{aligned}$$

Possui 3 ou 1 raízes positivas e, 2 ou nenhuma raízes negativas.

Por a_5 e a_4 serem nulos, pode-se concluir pela Regra da Lacuna que há raízes complexas conjugadas.

Pela Cota de Kojima, tem-se a seguinte série de fatores: $\{2, 1.8708, 1.8171, 0.8909, 0.9057\}$. A soma dos dois maiores fatores é 3.8708. Desta forma, pode-se afirmar que o módulo de toda e qualquer raiz de $p(x)$ é menor que 4.

a) VERDADEIRO

b) FALSO

c) VERDADEIRO

d) VERDADEIRO.

Exercício 6.3

$$\begin{aligned} q(x) &= (2x^5 + 6x^4 - 10x^3 - 30x^2 + 8x^1 + 24)x^2 \\ h(x) &= 2x^5 + 6x^4 - 10x^3 - 30x^2 + 8x^1 + 24 \\ S &= (+, +, -, -, +, +) \quad \Rightarrow \quad T = 2 \\ h(-x) &= -2x^5 + 6x^4 + 10x^3 - 30x^2 - 8x^1 + 24 \\ S' &= (-, +, +, -, -, +) \quad \Rightarrow \quad T' = 3 \end{aligned}$$

a) q(x) tem duas ou nenhuma raiz positiva

b) q(x) tem três ou uma raiz positiva

c) Não possui raízes complexas (verificar com a Regra da Lacuna)

d) As raízes são $\{0,\ 0,\ 1,\ 2,\ -1,\ -2,\ -3\}$ (tentar encontrar as cotas superior e inferior)

e) Não há raízes complexas.

Capítulo 7

Seção 7.7

Exercício 7.1

i) 4.4816890

ii) 5.0536689

Exercício 7.2

a) 20598

b) 20851

c) Pela Esquerda = 15479 Pela Direita = 26223.

Exercício 7.3 Considerando $\xi = 1$, por ser aonde $f''(\xi)$ possui maior módulo, tem-se que

$$\max\{E_{TTS}\} = 0.3372$$

Exercício 7.4

$$\begin{aligned} f(x) &= 2^4 - 3x^3 + 2x^2 - x + 1 - \sin\left(x - \frac{\pi}{2}\right) \\ f'(x) &= 8x^3 - 9x^2 + 4x - 1 - \cos\left(x - \frac{\pi}{2}\right) \\ f''(x) &= 24x^2 - 18x + 4 + \sin\left(x - \frac{\pi}{2}\right) \end{aligned}$$

Considerando $\xi = \pi$, por ser aonde $f''(\xi)$ possui maior módulo, e $h = \frac{\pi}{2}$:

$$\begin{aligned} \max\{E_{TTS}\} &= \frac{\left(\frac{\pi}{2}\right)^3}{12} \times \left(24\pi^2 - 18\pi + 4 + \sin\left(\frac{\pi}{2}\right)\right) \\ \max\{E_{TTS}\} &= 59.85562 \end{aligned}$$

Capítulo 8

Seção 8.10

Exercício 8.2 $i(3.8) = 0.0043$

Exercício 8.3 [Item a)] Partindo das equações diferenciais dadas no problema reescritas abaixo:

$$\begin{aligned} (M+m)\,\ddot{y} + ml\ddot{\theta} &= u \\ 2l\ddot{\theta} - 2g\theta + \ddot{y} &= 0 \end{aligned}$$

Primeiramente é preciso modificar as variáveis utilizadas, como sugerido no prrblema. As novas variáveis utilizadas são as seguintes:

$$\begin{cases} y &= x_1 \\ \dot{y} &= x_2 \\ \theta &= x_3 \\ \dot{\theta} &= x_4 \end{cases}$$

Com estas novas variáveis, as equações ficam da seguinte forma:

$$\begin{aligned} (M+m)\,\dot{x}_2 + ml\dot{x}_4 &= u \\ 2l\dot{x}_4 - 2gx_3 + \dot{x}_2 &= 0 \end{aligned}$$

Pode-se manipular estas duas equações de forma a obter um sistema da forma $\dot{x} = F(x,u)$. Isolando $\dot{x}_2$ da primeira equação e substituindo na segunda equação obtém-se a seguinte expressão:

$$\dot{x}_4 = \frac{2g\,(M+m)}{l\,(2M+m)}x_3 - \frac{u}{l\,(2M+m)}$$

Realizando o mesmo procedimento, agora isolando $\dot{x}_4$ da segunda equação e substituindo na primeira equação, têm-se o seguinte resultado:

$$\dot{x}_2 = \frac{-mg}{\left(M+\frac{m}{2}\right)}x_3 + \frac{u}{\left(M+\frac{m}{2}\right)}$$

Sabendo que $\dot{x}_1 = x2$ e $\dot{x}_3 = x_4$ pode-se montar o seguinte sistemas do tipo $\dot{x} = F(x,u)$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M+\frac{m}{2}} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{2g(M+m)}{l(2M+m)} & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{\left(M+\frac{m}{2}\right)} \\ 0 \\ \frac{-1}{l(2M+m)} \end{bmatrix} \cdot u$$

[Item b)] Substituindo os valores das variáveis do problema e da lei de controle $u = -kx$, pode-se simplificar as equações, chegando à um problema do tipo $\dot{x} = (A - bk)\, x$, dado por

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0,3085 & 0,9908 & 47,7577 & 10,7910 \\ 0 & 0 & 0 & 1 \\ -0,3085 & -0,9908 & -28,1577 & -10,7910 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \tag{C.1}$$

Este sistema de equações define o movimento do pêndulo invertido ao longo do tempo. Para calcular estas saídas aplica-se o método de Runge-Kutta para equações diferenciais do tipo $\dot{x} = F(x,t)$ com condição inicial $x(t_0) = x_0$ conhecida e h dado como o intervalo de integração .

Aplicando o método chega-se às seguintes expressões:

$$\begin{aligned} t_{k+1} &= t_k + h \\ x_{k+1} &= x_k + \frac{k_1 + k_2}{2} \\ k_1 &= h \cdot F(x_k, t_k) \\ k_2 &= h \cdot F(x_k + k_1, t_{k+1}) \end{aligned}$$

As condições iniciais do problema são conhecidas e dadas por:

$$\begin{aligned} t_0 &= 0 \\ h &= 0,1 \\ x_0 &= \begin{bmatrix} 0,2 & 0 & 0,5 & 0 \end{bmatrix}^T \\ \dot{x}(0) = \dot{x}_0 &= \begin{bmatrix} 0 & 23,9406 & 0 & -14,1406 \end{bmatrix}^T \end{aligned}$$

onde $\dot{x}_0$ foi calculado com o sistema de equações (C.1). Pode-se exemplificar a solução do problema calculando a primeira iteração. Para este caso têm-se os seguintes valores:

$$\begin{aligned} t_1 &= 0,1 \\ k_1 &= h \cdot \dot{x}_0 \\ &= \begin{bmatrix} 0 & 2,3941 & 0 & -1,4141 \end{bmatrix}^T \\ k_2 &= h \cdot \dot{x}_1 \\ &= h \cdot \dot{x}(x_0 + k_1, t_1) = \begin{bmatrix} 0,2394 & 1,1053 & -0,1414 & -0,1253 \end{bmatrix}^T \\ x(0,1) &= \begin{bmatrix} 0,3197 & 1,7497 & 0,4293 & -0,7697 \end{bmatrix}^T \end{aligned}$$

A tabela a seguir apresenta alguns pontos da trajetória de x.

	$x\,(0)$	$x\,(0,1)$	$\cdots$	$x\,(9,9)$	$x\,(10)$
x_1	$0,2$	$0,3197$	$\cdots$	$-0,0845$	$-0,0896$
x_2	$0,0$	$1,7497$	$\cdots$	$-0,0553$	$-0,0483$
x_3	$0,5$	$0,4293$	$\cdots$	$0,0037$	$0,0035$
x_4	$0,0$	$-0,7697$	$\cdots$	$-0,0022$	$-0,0021$

A seguir está um gráfico representando a evolução de $x\,(t)$ no intervalo de 0 à 10 segundos para as condições iniciais dadas.

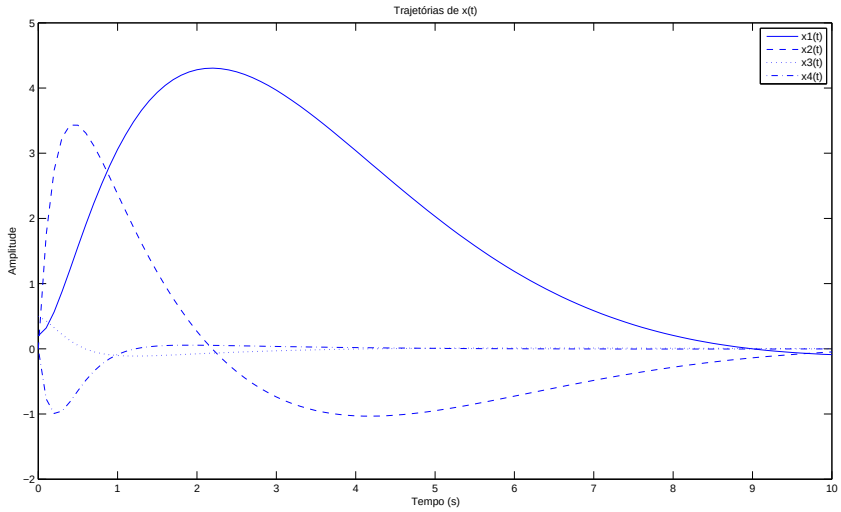


Figura C.1: Trajetória do sistema pêndulo invertido.

[Item c)] Modificando agora a condição inicial x_0 e aplicando o mesmo procedimento chega-se ao gráfico de $x\,(t)$ dado a seguir. Percebem-se variações em relação às amplitudes se comparadas ao gráfico anterior, porém apresentando comportamento semelhante ao longo da trajetória.

	$x\,(0)$	$x\,(0,1)$	$\cdots$	$x\,(9,9)$	$x\,(10)$
x_1	$-0,3$	$-0,1094$	$\cdots$	$-0,1201$	$-0,1286$
x_2	$0,0$	$2,7898$	$\cdots$	$-0,0908$	$-0,0800$
x_3	$0,8$	$0,6878$	$\cdots$	$0,0058$	$0,0054$
x_4	$0,0$	$-1,2218$	$\cdots$	$-0,0033$	$-0,0032$

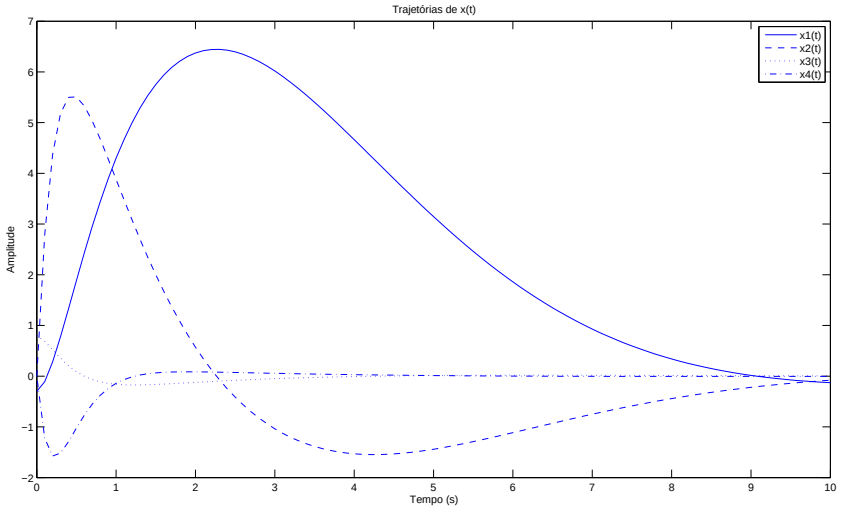


Figura C.2: Trajetória do sistema pêndulo invertido.

Exercício 8.4

$$\begin{cases} \dot{x_1} = \dot{V} = x_2 \\ \dot{x_2} = \ddot{V} = -\frac{4}{5}x_2 - \frac{1}{15}x_1 - 4e^{-2t} \end{cases}$$

a) $V(5) = -0.4365\, m/s \qquad \dot{V}(5) = -0.0306\, m/s^2$

b) $V(5) = -0.4343\, m/s \qquad \dot{V}(5) = -0.0307\, m/s^2$