CÁLCULO NUMÉRICO PARA ENGENHEIROS COM METODOLOGIAS NINJAS

Sérgio Mário Lins Galdino
Jornandes Dias da Silva
Cícero José da Silva
Willames de Albuquerque Soares
Juan Carlos Oliveira de Medeiros

EDITORA CONHECIMENTO LIVRE

Sérgio Mário Lins Galdino
Jornandes Dias da Silva
Cícero José da Silva
Willames de Albuquerque Soares
Juan Carlos Oliveira de Medeiros

Cálculo Numérico para Engenheiros com Metodologias Ninjas

1ª ed.

Piracanjuba-GO
Editora Conhecimento Livre
Piracanjuba-GO

1ª ed.

Dados Internacionais de Catalogação na Publicação (CIP)

Galdino, Sérgio Mário Lins
G171C Cálculo Numérico para Engenheiros com Metodologias Ninjas
/ Sérgio Mário Lins Galdino. Jornandes Dias da Silva. Cícero José da Silva. Willames de Albuquerque Soares. Juan Carlos Oliveira de Medeiros. – Piracanjuba-GO

Editora Conhecimento Livre, 2024

147 f.: il
DOI: 10.37423/2024.edcl999
ISBN: 978-65-5367-537-7
Modo de acesso: World Wide Web
Incluir Bibliografia

1. algoritmos-numéricos 2. análise-de-erros 3. modelagem-matemática I. Galdino, Sérgio Mário Lins II. Silva, Jornandes Dias da III. Silva, Cícero José da IV. Soares, Willames de Albuquerque V. Medeiros, Juan Carlos Oliveira de VI. Título

CDU: 510

https://doi.org/10.37423/2024.edcl999

EDITORA CONHECIMENTO LIVRE

Corpo Editorial

Dr. Francisco Odecio Sales

MSc Ezequiel Martins Ferreira

MSc Eliane Avelina de Azevedo Sampaio

MSc Carlos Eduardo De Oliveira Gontijo

Dr. Rodrigo Couto Santos

Dra. Milena Gaion Malosso

PHD Marcos Pereira Dos Santos

Editora Conhecimento Livre
Piracanjuba-GO
2024

10.37423/2024.edcl999

Agradecimentos

- Ao Diretor e Vice-Diretor da Escola Politécnica de Pernambuco Prof Dr. Alexandre Duarte Gusmão e Prof. Dr. Sérgio Campello Oliveira, pela compreensão e pelo suporte para realização deste.
- Ao Vice-Reitor da UPE e ex-Diretor da Escola Politécnica de Pernambuco Prof. Ms. José Roberto de Souza Cavalcanti, pela percepção e incentivo para efetuação deste.
- Aos amigos UPE-Poli (Universidade de Pernambuco-Escola Politécnica de Pernambuco) e da Unicap (Universidade Católica de Pernambuco), pelo incentivo e conhecimentos, os quais foram indispensáveis para construção, deste.
- A todos os colegas do Departamento Básico, pelo ânimo e amizade. Em especial, aos que contribuíram diretamente na concretização deste.
- Aos professores e amigos Prof. Cleto Bezerra de França, Prof. Roberto Lessa, Prof. Cláudio Maciel (poeta risadinha) e Prof. Emerson Alexandre de Oliveira Lima.
- Aos camaradas Prof. Fernando José Bertino de Figueirêdo e Prof.Antonio Carlos Miranda, pela nossa história na luta sindical classista e na politica revolucionária.

Dedicatória

Às Nossas Mães, Aos Nossos Pais, Filhos e Filhas, Esposas,Todos os Familiares e Amigos

Prefácio

Cálculo Numérico tem objetivo destacado o estudo de técnicas (ou métodos) numéricas para obter soluções de problemas que possam ser representados por modelos matemáticos. É uma importante ferramenta para a resolução de problemas práticos no engenharia e diversas áreas.

A resolução de modelos matemáticos é geralmente complexa, envolvendo fenômenos não-lineares, onde não é possível determinar soluções analíticas. Nestes casos, os métodos numéricos são ferramentas imprescindíveis nas soluções numéricas. Portanto, o cálculo numérico é fundamental na formação de profissionais das áreas de engenharias e ciências exatas.

Já se passaram quase 40 anos desde que o primeiro autor começou a ministrar cursos de cálculo numérico e introdutório de computação. Durante esse período, os métodos numéricos e os computadores obtiveram um papel proeminente no currículo de engenharia—particularmente nas partes iniciais. Muitas universidades agora oferecem cursos para calouros, alunos do segundo ano e do terceiro ano em algoritmos numéricos e introdução à computação. Muitos outros cursos estão integrando problemas orientados por computador em todos os níveis do currículo. Assim, os estudantes de engenharia devem receber uma introdução forte e precoce aos métodos numéricos. Consequentemente, procuramos utilizar características que tornam o conteúdo acessível tanto para estudantes de níveis iniciais quanto avançados. Essas características incluem:

Orientação para Problemas. Os estudantes de engenharia aprendem melhor quando são motivados por problemas práticos. Apresentamos os métodos numéricos na perspectiva da solução de problemas.

Metodologia Ninja. Orientada para o aluno, desenvolvemos uma série de recursos para tornar este livro o mais acessível possível aos estudantes. Isso inclui a organização geral, além do uso extensivo de exemplos resolvidos (com calculadoras e/ou computadores) na "Metodologia Ninja"(combinando solução rápida e menos propenso a erros). Também tentamos manter nossas explicações diretas e orientadas de forma prática.

Ferramentas Computacionais. Capacitamos nossos estudantes a ajudá-los na resolução numérica de problemas através de Ferramentas Computacionais com uso de *softwares* (Python, Octave, MATLAB e Mathcad, Maple, Mathematica - *Stand-Alone* ou *Online*). No entanto, também estimulamos os estudantes a desenvolver programas simples e bem estruturados para expandir as capacidades básicas desses ambientes. A atual fuga da programação de computadores representa uma certa "simplificação" do currículo de engenharia. Os engenheiros não satisfeitos em serem limitados por ferramentas terão que escrever código.

Esperamos que este curso forneça habilidades para: compreender como os números são representados nas calculadoras e computadores para realizar computações numéricas; conhecer e aplicar os principais métodos numéricos para a solução de problemas práticos; estimar e analisar os erros obtidos; e propor soluções para minimizá-los ou mesmo, quando possível, eliminá-los. A participação nas atividades e em cada aula é essencial para que você possa tirar o maior proveito da disciplina.

Finalmente, registro que em 1988 ministrei aula de Cálculo Numérico na Unicap (Universidade Católica de Pernambuco) ao Prof. Cícero José da Silva - um aluno destacado. Este foi o primeiro ano que ministrei aulas de Cálculo Numérico. No ano seguinte já era meu colega de trabalho na Unicap. Em 2009 cheguei na Poli-UPE e encontrei novamente o Prof. Cícero que agora lidera o *Grupo de Matemática e áreas afins* na UPE-Poli.

Nota

Escrever um texto desta natureza demanda tempo e nos causa uma certo cuidado adicional pela equipe, pois se teme cometer os erros que ensinamos evitar. Por este motivo, sugestões, corrreções, comentários, antecipadamente agradecemos, devem ser enviados para um dos endereços:

galdino.sergio@gmail.com
jornandesdias@poli.br
juca@ufc.br
cjs@poli.br
was@poli.br

Recife, 01 de agosto de 2024

Conteúdo

Capítulo - Erros em computações numéricas

1.1 Sistemas de numeração

O sistema de numeração romana (algarismos romanos ou números romanos) desenvolveu-se na Roma Antiga, e foi utilizado em todo o Império Romano.

Número romano	Nome	Valor
I	unus	1 (um)
V	quinque	5 (cinco)
X	decem	10 (dez)
L	quinquaginta	50 (cinquenta)
C	centum	100 (cem)
D	quingenti	500 (quinhentos)
M	mille	1000 (mil)

Os números são escritos através de algarismos, começando do algarismo de maior valor e seguindo a seguinte regra:

- Algarismos de menor ou igual valor à direita são somados ao algarismo de maior valor;
- Algarismos de menor valor à esquerda são subtraídos do algarismo de maior valor.

Exemplo 1.1

XI $= 10 + 1 = 11$

XC $= 100 - 10 = 90$

MCMLVII $= 1000 + (1000 - 100 = 900) + 50 + 5 + 1 + 1 = 1957$

Características dos Números Romanos

Zero. Os romanos desconheciam o zero o início da contagem é pelo valor I $= 1$. O zero ocorreu originalmente em três povos: babilônios, hindus e maias. Na Europa, a definição do zero ocorreu na Idade Média, após a adoção dos algarismos arábicos, difundidos po Leonardo Fibonacci.

Sistema de numeração não posicional. O simbolo representa sempre a mesma grandeza - como é observado no números romanos é chamado de sistema de numeração não-posicional.

Números Decimais

Números decimais são numerais que se usa uma vírgula (ou ponto), indicando que o algarismo a seguir pertence à ordem das décimas, ou casas decimais. Todos os números decimais finitos ou infinitos e periódicos podem ser escritos na forma de fração (são os números racionais).

Exemplo 1.2

1957 e 23.5332

$\sqrt{2}$ e π não são números decimais

$\frac{1}{3} = 0.33333\cdots = 0.\overline{3}$ e $\frac{25}{37} = 0.\overline{675}\cdots$

$\frac{\pi}{4} \approx 0.78539816$ o lado esquerdo é uma constante simbólica de um número real divido por 4, o lado esquerdo é um número decimal aproximado

Sistema de numeração posicional

É um modo de representação numérica na qual o valor de cada algarismo depende da sua posição relativa na composição do número.
Um número x pode ser representado num sistema de base b conforme o polinômio:

$$x = d_{n-1}b^{n-1} + d_{n-2}b^{n-2} + \cdots + d_1b^1 + d_0b^0 + \\ +d_{-1}b^{-1} + d_{-2}b^{-2} + \cdots + d_{-m}b^{-m}$$

Onde n é a quantidade de dígitos inteiros e m a quantidade de dígitos fracionários
Usualmente o número x é representado em base b pelos algarismos concatenados da seguinte forma:

$$x = (d_{n-1}d_{n-2}\cdots d_1d_0, d_{-1} + d_{-2}\cdots d_{-m})_b$$

Quando $b = 10$, a indicação da base é usualmente suprimida.

O sistema padrão é o decimal, a base é 10 e utiliza os algarismos de 0 a 9.
Nos sistemas de informação são bastante utilizados: o binário de base 2, que tem como algarismos 0 e 1 e o hexadecimal de base 16 e utiliza os dígitos de 0 a 9 e as letras de A a F. Assim, o número $x = 42$ pode ser escrito em binário como $x = 101010_2$ e em hexadecimal como $x = 2\text{A}_{16}$.

A base é o número de algarismos diferentes que podem ser utilizados para representar os números.

Exemplo 1.3

$123.532 = 1 \cdot 10^2 + 2 \cdot 10^1 + 3 \cdot 10^0 + 3 \cdot 10^0 + 5 \cdot 10^{-1} + 3 \cdot 10^{-2} + 2 \cdot 10^{-3}$
$2\text{A}_{16} = 2 \cdot 16^1 + 10 \cdot 16^0 = 2 \cdot 16 + 10 = 42$
$101010_2 = 1 \cdot 2^5 + 1 \cdot 2^3 + 1 \cdot 2^1 = 32 + 8 + 2 = 42$

O sistema de numeração decimal é o mais usado pelo homem nos dias de hoje. O número 10 tem papel fundamental, é chamado de base do sistema. Os símbolos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, são usados para representar qualquer grandeza. O fato de o sistema decimal ser largamente utilizado tem evidentemente razões históricas, pois na realidade qualquer número inteiro maior que 1 poderia ter sido escolhido. De fato, no mundo dos computadores digitais o sistema binário é o utilizado. O número 2 é a base do sistema e os símbolos 0 e 1 servem para representar uma grandeza qualquer. Ao lado do sistema binário, os sistemas octal, hexadecimal, base 8 e 16 respectivamente, são também utilizados. Isto ocorre pelo fato de que cada símbolo octal e hexadecimal representa um equivalente a três e quatro símbolos no sistema binário e vice-versa. A maior parte dos sistemas de computação atuais *hardware* e *software* usam a representação binária internamente, embora muitos computadores antigos, tal como o ENIAC ouIBM 650, usaram o sistema decimal internamente.

1.2 Computações Numéricas

A aritmética das máquinas digitais (computadores, calculadoras, ...) não é a mesma que é usada em cursos de cálculo ou álgebra. Assume-se como afirmações verdadeiras que $2 + 2 = 4$, $2^2 = 4$, e $(\sqrt{2})^2 = 2$. Na aritmética padrão das máquinas digitais as duas primeiras são verdadeiras mas a terceira não. Para entender poque isto é verdadeiro deve-se explorar o mundo da aritmética de precisão finita utilizada por máquinas digitais.

1.2.1 Conversão

Dado um número x representado na base N, isto é, na N-representação, e nós queremos saber como representa-lo na base M, isso é, na M-representação. Temos então a equação:

$$\begin{aligned} x &= a_mN^m + \cdots + a_1N^1 + a_0N^0 + a_{-1}N^{-1} + \cdots + a_{-n}N^{-n} \\ &= b_jM^j + \cdots + b_1M^1 + b_0M^0 + b_{-1}M^{-1} + \cdots + b_{-k}M^{-k} \end{aligned} \tag{1.1}$$

onde os coeficientes $a_m,\ a_{m-1},\ \cdots,\ a_1,\ a_0,\ a_{-1},\ \cdots,\ a_{n-1},\ a_{-n}$ são conhecidos e os coeficientes $b_j,\ b_{j-1},\ \cdots,\ b_1,\ b_0,\ b_{-1},\ \cdots,\ b_{k-}$ devem ser determinados. Observe que $b_j,\ b_{j-1},\ \cdots,\ b_1,\ b_0,\ b_{-1},\ \cdots,\ b_{k-1},\ b_{-k}$ devem ser expressos com símbolos de dígitos da N-representação. Para realizar a conversão dividiremos x em uma parte inteira i e uma parte fracionária f.

Nós temos $i = b_j M^j + \cdots + b_1 M^1 + b_0 M^0$, e dividindo i por M nós obtemos um quociente q_1 e um resto $r_1 = b_0$. Continuando, dividiremos q_1 por M, nós conseguiremos q_2 e o resto $r_2 = b_1$, e, obviamente, $b_0,\ b_1,\ b_2,\ \cdots$ são os restos consecutivos quando i é dividido repetitivamente por M. De forma semelhante nós encontramos a parte fracionária como as partes inteiras consecutivas quando f é multiplicado repetitivamente por M e a parte inteira é removida. Os cálculos devem ser feitos na N-representação e M deve ser também dado nesta representação.

Exemplo 1.4

Conversão o número decimal $261,359$ para a representação binária.

Conversão: Decimal para binário

Inteiro: Divisão sucessiva do número decimal por 2

$$
\begin{array}{rl}
261 & 2 \\
\cline{2-2}
1 & 130 \quad 2 \\
\cline{3-3}
0 & 65 \quad 2 \\
\cline{4-4}
1 & 32 \quad 2 \\
\cline{5-5}
0 & 16 \quad 2 \\
\cline{6-6}
0 & 8 \quad 2 \\
\cline{7-7}
0 & 4 \quad 2 \\
\cline{8-8}
0 & 2 \quad 2 \\
\cline{9-9}
0 & 1
\end{array}
$$

O número inteiro binário é obtido através dos restos das divisões escritos na ordem inversa da sua obtenção. Então $261_{10} = 1.0000.0101_2$.

Fração: Multiplicação sucessiva da fração decimal por 2

Multiplicação	Sobra
0,359x2 = 0,718	0
0,718x2 = 1,436	1
0,436x2 = 0,872	0
0,872x2 = 1,774	1
0,774x2 = 1,488	1
0,488x2 = 0,976	0
0,976x2 = 1,952	1
0,952x2 = 1,904	1
0,904x2 = 1,808	1
$\vdots \quad \vdots \quad \vdots$	$\vdots$

A fração binária é obtida através das sobras, parte inteira, das multiplicações escritas na ordem direta de sus obtenção. Então $0,359_{10} = 0,0101.1011.1\cdots_2$.

Somando-se as partes inteiras e fracionárias dos binários obtidos têm-se

$$261,359_{10} = 1.0000.0101,0101.1011.1\cdots_2$$

Exemplo 1.5

Conversão o número decimal $261,359$ para a representação ternária.

Conversão: Decimal para ternário

Inteiro: Divisão sucessiva do número decimal por 3

$$\begin{array}{rl|l} 261 & 3 & \\ \cline{2-3} 0 & 87 & 3 \\ \cline{3-3} & 0 & 29 \quad 3 \\ \cline{4-4} & & 2 \quad 9 \quad 3 \\ \cline{5-5} & & & 0 \quad 3 \quad 3 \\ \cline{6-6} & & & & 0 \quad 1 \end{array}$$

O número inteiro ternário é obtido através dos restos das divisões escritos na ordem inversa da sua obtenção. Então $261_{10} = 100.200_3$.

Fração: Multiplicação sucessiva da fração decimal por 3

Multiplicação	Sobra
0,359x3 = 1,077	1
0,077x3 = 0,231	0
0,231x3 = 0,693	0
0,693x3 = 2,079	2
0,079x3 = 0,237	0
0,237x3 = 0,711	0
0,711x3 = 2,133	2
0,133x3 = 0,399	0
0,399x3 = 1,197	1
$\vdots \quad \vdots \quad \vdots$	$\vdots$

A fração ternária é obtida através das sobras, parte inteira, das multiplicações escritas na ordem direta de sus obtenção. Então $0,359_{10} = 0,100.200.201\cdots_3$. Somando-se as partes inteiras e fracionárias dos binários obtidos têm-se

$$261,359_{10} = 100.200,100.200.201\cdots_3$$

Exemplo 1.6

Conversão o número decimal $261,359$ para a representação hexadecimal.

Conversão: Decimal para hexadecimal

Inteira: Divisão sucessiva do número decimal por 16

$$\begin{array}{rl|l} 261 & 16 & \\ \cline{2-3} 5 & 16 & 16 \\ \cline{3-3} & 0 & 1 \end{array}$$

O número inteiro hexadecimal é obtido através dos restos das divisões escritos na ordem inversa da sua obtenção. Então $261_{10} = 105_{16}$.

Fração: Multiplicação sucessiva da fração decimal por 16

Multiplicação	Sobra
0,359x16 = 5,744	5
0,744x16 = 11,904	11
0,904x16 = 14,464	14
0,464x16 = 7,424	7
0,424x16 = 6,784	6
0,784x16 = 12,544	12
0,544x16 = 8,704	8
0,704x16 = 11,264	11
0,264x16 = 4,224	4
⋮ ⋮ ⋮	⋮

A fração hexadecimal é obtida através das sobras, parte inteira, das multiplicações escritas na ordem direta de sus obtenção. Então

$$0,359_{10} = 0,5\ 11\ 14\ 7\ 6\ 12\ 8\ 11\ 4\ \cdots_{16}$$

ou utilizando-se os símbolos hexadecimais (Tabela 1.1)

$$0,359_{10} = 0,5\text{BE.}76\text{C.}8\text{B}4\cdots_{16}$$

Somando-se as partes inteiras e fracionárias dos hexadecimais obtidos têm-se

$$261,359_{10} = 105,5\text{BE.}76\text{C.}8\text{B}4\cdots_{16}$$

Tabela 1.1: Símbolos Hexadecimais.

Grandeza decimal	Símbolo hexadecimal
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A
11	B
12	C
13	D
14	E
15	F

Conv10.sce

Este roteiro Scilab realiza a conversão de inteiros na base 10 para base qualquer.

```
1.  // Conversão de inteiros da base 10 para uma base qualquer
2.  function y=conv10(x,b)
3.  y=''
```

```
4.  while x > 0
5.     cx=modulo(x,b)
6.     x=int(x/b)
7.     y=string(cx)+'.'+y
8.  end
9.  endfunction
```

Ao executar o roteiro **Conv10.sce** pode-se realizar as conversões. Aqui estão os resultados:

```
-->conv10(261,2)
 ans  =

 1.0.0.0.0.0.1.0.1.

-->

-->conv10(261,3)
 ans  =

 1.0.0.2.0.0.

-->

-->conv10(261,16)
 ans  =

 1.0.5.

-->
```

ConvF10.sce

Este roteiro Scilab realiza a conversão de fracionários na base 10 para base qualquer.

```
1.  // Conversão de fracionários da base 10 para uma base qualquer
2.  function y=convF10(x,b,n)
3.  y='0,'
4.  i=0
5.  while i < n
6.     p=x*b
7.     cx=int(p)
8.     y=y+string(cx)+'.'
9.     x=p-cx
10.    i=i+1
11. end
12. endfunction
```

Ao executar o roteiro **ConvF10.sce** pode-se realizar as conversões. Aqui estão os resultados:

```
-->convF10(0.359,2,9)
 ans  =

 0,0.1.0.1.1.0.1.1.1.

-->

-->convF10(0.359,3,9)
 ans  =

 0,1.0.0.2.0.0.2.0.1.

-->

-->convF10(0.359,16,9)
 ans  =

 0,5.11.14.7.6.12.8.11.4.
```

```
-->
```

Conv10X.sce

Este roteiro Scilab realiza a conversão de um número na base 10 para base qualquer (combina **Conv10.sce** e **ConvF10.sce**) .

```
1.  // Conversão da base 10 para uma base qualquer: Inteiro + Fração
2.  function y=conv10X(xx,b,n)
3.  x=int(xx)
4.  yi=''
5.  while x > 0
6.     cx=modulo(x,b)
7.     x=int(x/b)
8.     yi=string(cx)+'.'+yi
9.  end
10. y=yi
11. x=xx-int(xx)
12. yf=','
13. i=0
14. while i < n
15.     p=x*b
16.     cx=int(p)
17.     yf=yf+string(cx)+'.'
18.     x=p-cx
19.     i=i+1
20. end
21. y=y+yf
22. endfunction
```

Ao executar o roteiro **Conv10X.sce** pode-se realizar as conversões. Aqui estão os resultados:

```
-->conv10X(261,2,0)
 ans  =

 1.0.0.0.0.0.1.0.1.,

-->

-->conv10X(0.359,2,9)
 ans  =

 ,0.1.0.1.1.0.1.1.1.

-->

-->conv10X(261.359,2,9)
 ans  =

 1.0.0.0.0.0.1.0.1.,0.1.0.1.1.0.1.1.1.

-->

-->conv10X(261,16,0)
 ans  =

 1.0.5.,

-->

-->conv10X(0.359,16,9)
 ans  =

 ,5.11.14.7.6.12.8.11.4.

-->

-->conv10X(261.359,16,9)
 ans  =
```

```
1.0.5.,5.11.14.7.6.12.8.11.4.

-->
```

1.2.2 Conversão com uso de tabela: Octal ⇄ Binário

A conversão de números de base 8 ($= 2^3$) para a base 2 é feita convertendo cada dígito no seu equivalente binário de 3 bits.

Para fazer a conversão de números binários para octais, utiliza-se a mesma tabela de conversão utilizada para converter números octais em binários. O agrupamento é relizado na direção esquerda para inteiros e na direção direita para frações. Caso o binário não tenha grupos regulares de 3 bits, começando podem ser adicionados até 2 0s à esquerda da parte inteira ou até 2 0s à direita da parte fracionária.

A tabela abaixo mostra o equivalente binário de cada dígito do sistema octal:

Dígito octal	Binário Equivalente
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Exemplo 1.7 *Conversão Binário → Octal*

Exemplos:

a) $11010111110_2 =$

$\underbrace{011}_{3}\ \underbrace{010}_{2}\ \underbrace{111}_{7}\ \underbrace{110}_{6}{}_2 = 3276_8$

b) $0{,}1101001_2 =$

$0{,}\underbrace{110}_{6}\ \underbrace{100}_{4}\ \underbrace{100}_{4}{}_2 = 0{,}644_8$

c) $10101{,}10101_2 =$

$\underbrace{010}_{2}\ \underbrace{101}_{5}{,}\ \underbrace{101}_{5}\ \underbrace{010}_{2}{}_2 = 25{,}52_8$

Exemplo 1.8

Conversão Octal → Binário a) $3752_8 = \underbrace{3}_{011}\ \underbrace{7}_{111}\ \underbrace{5}_{101}\ \underbrace{2}_{010}{}_8 = 11\ 111\ 101\ 010_2$

b) $0{,}426_8 = 0{,}\underbrace{4}_{100}\ \underbrace{2}_{010}\ \underbrace{6}_{110}{}_8 = 0{,}100\ 010\ 11_2$

c) $63{,}741_8 = \underbrace{6}_{110}\ \underbrace{3}_{011}{,}\ \underbrace{7}_{111}\ \underbrace{4}_{100}\ \underbrace{1}_{001}{}_8$

$= 110\ 011{,}111\ 100\ 001_2$

1.2.3 Conversão com uso de tabela: Hexadecimal ⇄ Binário

A conversão de números de base 16 ($= 2^4$) para a base 2 é feita convertendo cada dígito no seu equivalente binário de 4 bits.

Para fazer a conversão de números binários para hexadecimais, utiliza-se a mesma tabela de conversão utilizada para

converter números hexadecimais em binários. O agrupamento é relizado na direção esquerda para inteiros e na direção direita para frações.
Caso o binário não tenha grupos regulares de 3 bits, começando podem ser adicionados até 3 0s à esquerda da parte inteira ou até 3 0s à direita da parte fracionária.

As tabelas abaixo mostram o equivalente binário de cada dígito do sistema hexadecimal:

Dígito hexa-decimal	Binário Equivalente
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Dígito hexa-decimal	Binário Equivalente
8	1000
9	1001
A	1010
B	1011
C	1100
D	1101
E	1110
F	1111

Exemplo 1.9

Conversão Binário → Hexadecimal

a) $11010111110001_2 = \underbrace{0011}_{3}\underbrace{0101}_{5}\ \underbrace{1111}_{F}\ \underbrace{0001}_{1}{}_2$
$= 35F1_{16}$

b) $0{,}0110001000111_2 = 0,\underbrace{0110.}_{6}\ \underbrace{0010.}_{2}\ \underbrace{0011}_{3}\ \underbrace{1000}_{8}{}_2$
$= 0{,}6238_{16}$

c) $110000011{,}11010111_2 = \underbrace{0001}_{1}\ \underbrace{1000}_{8}\ \underbrace{0011}_{3},\underbrace{1101}_{D}\ \underbrace{0111}_{7}{}_2$
$= 183{,}D7_{16}$

Exemplo 1.10

Conversão Hexadecimal → Binário

a) $3A5F_{16} = \underbrace{3}_{0011}\ \underbrace{A}_{1010}\ \underbrace{5}_{0101}\ \underbrace{F}_{1111}{}_{16} = 11\ 1010\ 0101\ 1111_2$

b) $0{,}4C6_{16} = 0,\underbrace{4}_{0100}\ \underbrace{C}_{1100}\ \underbrace{6}_{0110}{}_{16} = 0{,}0100\ 1100\ 0110_2$

c) $A6{,}97E_{16} = \underbrace{A}_{1010}\ \underbrace{6}_{0110},\underbrace{9}_{1001}\ \underbrace{7}_{0111}\ \underbrace{E}_{1110}{}_{16}$
$= 1010\ 0110{,}1001\ 0111\ 111_2$

1.2.4 Conversão de uma base qualquer para base 10

Para obter-se o número decimal equivalente a um número escrito em qualquer base é só multiplicar cada dígito por sua potência:

Exemplo 1.11

$261,359_{10} \cong 1.0000.0101,0101.1011.1_2$

Parte inteira

$$1 \cdot 2^8 + 0 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 = 261$$

Parte fracionária

$$0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 0 \cdot 2^{-6} + 1 \cdot 2^{-7} + 1 \cdot 2^{-8} + 1 \cdot 2^{-9} \cong 0,3574$$

Erro

$|261,359 - 261,3574| = 0,0016$

Exemplo 1.12

$261,359_{10} \cong 100.200,100.200.201_3$

Parte inteira

$$1 \cdot 3^5 + 0 \cdot 3^4 + 0 \cdot 3^3 + 2 \cdot 3^2 + 0 \cdot 3^1 + 0 \cdot 3^0 = 261$$

Parte fracionária

$$1 \cdot 3^{-1} + 0 \cdot 3^{-2} + 0 \cdot 3^{-3} + 2 \cdot 3^{-4} + 0 \cdot 3^{-5} + 0 \cdot 3^{-6} + 2 \cdot 3^{-7} + 0 \cdot 3^{-8} + 1 \cdot 3^{-9} \cong 0,35899$$

Erro

$|261,359 - 261,35899| = 0,00001$

Exemplo 1.13

$261,359_{10} \cong 105,5\ 11\ 14\ 7\ 6\ 12\ 8\ 11\ 4_{16}$ ou seja,

$261,359_{10} \cong 105,5\text{BE}.76\text{C}.8\text{B}4_{16}$

Parte inteira

$$1 \cdot 16^2 + 0 \cdot 16^1 + 5 \cdot 16^0 = 261$$

Parte fracionária

$$\frac{5}{16^1} + \frac{11}{16^2} + \frac{14}{16^3} + \frac{7}{16^4} + \frac{6}{16^5} + \frac{12}{16^6} + \frac{8}{16^7} + \frac{11}{16^8} + \frac{4}{16^9}$$

$$\cong 0,35899999999674$$

Erro

$$|261,359 - 0,35899999999674| \cong 3,24 \times 10^{-12}$$

1.2.5 Representação de um número inteiro

A representação de um **número inteiro** num computador qualquer que trabalha internamente com uma base fixa $\beta \geq 2$ (um número inteiro); sendo geralmente escolhido como uma potência de 2.

Dado um número inteiro $n \geq 0$, ele possui uma única representação,

$$\begin{aligned} n &= \pm(n_k n_{k-1} \cdots n_1 n_0) \\ &= \pm(n_k \beta^k + n_{k-1}\beta^{k-1} + \cdots + n_1\beta^1 + n_0\beta^0), \end{aligned} \tag{1.2}$$

onde os $n_i,\ i = 0,\ 1,\ \cdots,\ k$ são inteiros satisfazendo $0 \leq n_i < \beta$ e $n_k \neq 0$.

Por exemplo, na base $\beta = 10$, o número 1957 é representado por:

$$1957 = 1 \times 10^3 + 9 \times 10^2 + 5 \times 10^1 + 7 \times 10^0$$

e é armazenado como $n_3 n_2 n_1 n_0$.

1.2.6 Representação de um número real

A representação de um número real no computador pode ser feita de duas maneiras:

1.2.6.1 a) Representação do ponto fixo

Este foi o sistema usado, no passado, por muitos computadores. Sendo ainda usado por ser simples em aplicações de microprocessadores. Um número real, $x \neq 0$, ele será representado em ponto fixo por:

$$x = \pm \sum_{i=k}^{-l} x_i \beta^i \tag{1.3}$$

onde k e l são o comprimento da parte inteira e fracionária do número x, respectivamente.

Por exemplo, na base $\beta = 10$, o número 1957.325 é representado por:

$$\begin{aligned} 1957.325 &= + \sum_{i=3}^{-3} x_i \beta^i \\ &= 1 \times 10^3 + 9 \times 10^2 + 5 \times 10^1 + 7 \times 10^0 + 3 \times 10^{-1} + 2 \times 10^{-2} + 5 \times 10^{-3} \end{aligned}$$

e é armazenado como $x_3 x_2 x_1 x_0 . x_{-1} x_{-2} x_{-3}$.

1.2.6.2 b) Representação do ponto-flutuante

Em geral, um número x na representação do ponto-flutuante tem a forma seguinte:

$$x = \pm m \cdot \beta^k \tag{1.4}$$

onde

$m =$ mantissa, um número fracionário em ponto fixo, isto é, $m = \sum_{i=1}^{t} m_{-i}\beta^{-i}$ (onde $i = 1,\ 2,\ \cdots t$, se $x \neq 0$, então $0 < m < 1$ e t é a quantidade de dígitos significativos ou precisão do sistema);

$\beta =$ base, 2 se o sistemas de numeração for binário, 10 se o sistema de $k =$ expoente, um inteiro $(k_{inf} \leq k_{sup})$.

Nota $m_{-1} \neq 0$ *ou* $\frac{1}{\beta} \leq m < 1$ *(significa que devemos ter um digito não nulo após a virgula) caracteriza o sistema de números em ponto flutuante normalizado.*

Nota *Em números binários de ponto flutuante IEEE 754, valores zero são representados pelo expoente polarizado e significando que ambos sendo zero. O zero negativo tem o bit de sinal definido para um. Um pode obter zero negativo como o resultado de determinados cálculos, por exemplo, como resultado de subfluxo aritmético em um número negativo, ou* -1.0×0.0*, ou simplesmente como* -0.0*.*

Na codificação de ponto flutuante decimal IEEE 754, um zero negativo é representado por um expoente que é qualquer expoente válido no intervalo para a codificação, o verdadeiro significando sendo zero, e o bit de sinal sendo um.

Exemplo 1.14

Escrever o número $N = -19.2 \cdot 10^{-8}$ em ponto flutuante na forma normalizada.

Reescrevendo o número para a forma $N = -0.192 \cdot 10^{-6}$, o número fica na representação do ponto-flutuante, o expoente é igual a -6, a mantissa é igual a -0.192 e a base é 10.

Escrevendo agora os números: $x_1 = 0.53; x_2 = -8.472; x_3 = 0.0913; x_4 = 35391.3$ e $x_5 = 0.0004$, onde todos estão na base $\beta = 10$, em ponto flutuante na forma normalizada:

$$\begin{aligned} x_1 &= 0.53 = 0.53 \times 10^0, \\ x_2 &= -8.472 = -0.8472 \times 10^1, \\ x_3 &= 0.0913 = 0.913 \times 10^{-1}, \\ x_4 &= 35391.3 = 0.353913 \times 10^5, \\ x_5 &= 0.0004 = 0.4 \times 10^{-3}. \end{aligned}$$

Para representarmos um sistema de números em ponto flutuante normalizado, na base β, com t dígitos significativos e com limites do expoente k_{inf} e k_{sup}, usa-se a notação: $\mathbb{F}_N(\beta, t, k_{inf}, k_{sup})$.

Um número em $\mathbb{F}_N(\beta, t, k_{inf}, k_{sup})$ será representado por:

$$\pm 0.m_{-1}m_{-2}m_{-3}\cdots m_{-t} \times \beta^k \tag{1.5}$$

onde $m_{-1} \neq 0$ e $k_{inf} \leq k \leq k_{sup}$.

Exemplo 1.15

Considere o sistema $\mathbb{F}_N(10, 3, -2, 2)$. Represente nesse sistema os números do exemplo anterior.

Solução: Os número serão representado por $\pm 0.m_{-1}m_{-2}m_{-3} \times \beta^{-k}$, onde $-2 \leq k \leq 2$. Então:

Normalização	$\mathbb{F}_N(10, 3, -2, 2)$
$x_1 = 0.53 = 0.53 \times 10^0$	0.530×10^0
$x_2 = -8.472 = -0.8472 \times 10^1$	-0.847×10^1
$x_3 = 0.0913 = 0.913 \times 10^{-1}$	0.913×10^{-1}
$x_4 = 35391.3 = 0.353913 \times 10^5$	(0.353×10^5)
$x_5 = 0.0004 = 0.4 \times 10^{-3}$	(0.400×10^{-3})

Observe que os números $x_4 = 35391.3$ e $x_5 = 0.0004$ não podem ser representados no sistema. De fato, o número $35391.3 = 0.353913 \times 10^5$ tem o expoente maior que 2, causando **overflow**, por outro lado $0.0004 = 0.4 \times 10^{-3}$ e assim o expoente é menor que -2 causando **underflow**.

Exemplo 1.16

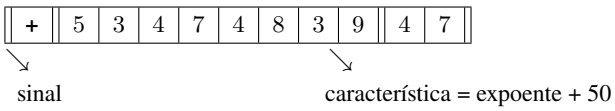
Diferença entre dois números na aritmética em ponto flutuante normalizada.

$$\begin{array}{r} 0.27143247 \cdot 10^7 \\ -0.27072236 \cdot 10^7 \\ \hline 0.00071011 \cdot 10^7 \end{array}$$

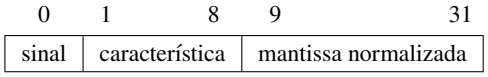
Vemos que a diferença entre estes dois números ponto-flutuante normalizados, resulta num número em ponto-flutuante não normalizado. Podemos, entretanto, normaliza-lo se deslocarmos o ponto três lugares à direita e somar -3 ao expoente, obtendo-se $0.71011000 \cdot 10^4$ normalizado.

1.2.7 Armazenamento na memória

Para começar vamos representar o número 0.00053474839 num computador decimal. A notação ponto-flutuante normalizada deste número é $0.53474839 \cdot 10^{-3}$. Par evitar expoente negativo, nós adicionamos, arbitrariamente, 50 (deslocamento) ao expoente e o número agora é $0.53474839 \cdot 10^{47}$. O expoente somado a uma constante arbitrária é chamado de característica. O número pode ser representado, unicamente, através da normalização da notação ponto-flutuante, na memória do computador utilizando o esquema de representação seguinte:



Deve ser observado que a característica coloca o expoente limitado a expressão seguinte: $-50 \leq k \leq 49$. O número tem o máximo de oito dígitos de precisão e a representação falha quando temos números muito grande ou muito pequeno. De modo análogo, um número binário na representação do ponto-flutuante também pode ser armazenado na memória de um computador digital. Uma palavra armazenada tendo um bit de "sinal"e 31 bits regulares pode representar um número binário ponto-flutuante na forma seguinte:

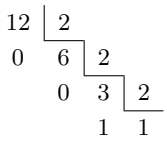


onde

sinal = sinal do número codificado, 0 se positivo e 1 se negativo;
característica = 127 + expoente (resultado escrito em binário);
mantissa = fração binária normalizada.

Exemplo 1.17

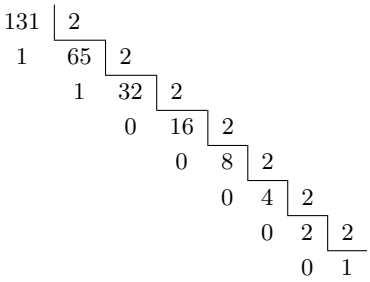
Represente o número 12.625 numa palavra de 32 bits conforme esquema de representação acima.



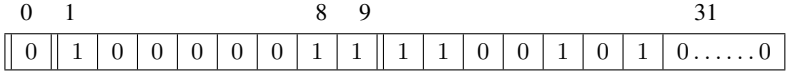
Multiplicação	Sobra
0.625x2 = 1.25	1
0.25 x2 = 0.50	0
0.5 x2 = 1.0	1

$12.625_{10} = 1100.101_2 = 0.1100101 \times 2^4$

Ajustando a característica: $4 + 127 = 131$



$131_{10} = 1000\ 0011_2$



1.2.8 Aritmética do ponto-flutuante

Os princípios das operações aritméticas básicas de um computador serão discutidos agora. Para isto iremos considerar que estamos trabalhando num computador decimal com uma palavra de 10 dígitos de comprimento. Princípios semelhantes são utilizados em computadores binários (digitais).

Operação Adição ($+$ ou $-$)

Na adição (soma ou subtração) de dois números o computador examina a característica ajustada dos dois números. Os seguintes casos são possíveis:

1- Características iguais: Adiciona-se as mantissas e mantém-se a característica.

$$\begin{array}{r} 32109876\ \underline{54} \\ +12340123\ \underline{54} \\ \hline 44449999\ \underline{54} \end{array}$$

2- Quando existe estouro (overflow) na adição das mantissas: O resultado será ajustado.

$$\begin{array}{r} 51319212\ \underline{55} \\ +98756431\ \underline{55} \\ \hline \underline{1}50075643\ \underline{55} \end{array}$$

$\searrow$ estouro $\quad \searrow$ característica

Resulta em:

$$15007564\ \underline{56}$$

$\searrow$ característica ajustada

3- Características distintas: Mantém-se a de maior módulo e ajusta-se a de menor valor.

$$\begin{array}{r} 31411122\ \underline{55} \\ +12344321\ \underline{53} \\ \hline \end{array} \quad \longrightarrow \quad \begin{array}{r} 31411122\ \underline{55} \\ +00123443\ \underline{55} \\ \hline 31534565\ \underline{55} \end{array}$$

4- Resultado com zero, ou zeros, à esquerda: Normaliza-se o resultado.

$$\begin{array}{r} 34122222\ \underline{73} \\ -34000122\ \underline{73} \\ \hline 00122100\ \underline{73} \end{array} \quad \text{resulta em:} \quad 12210000\ \underline{71}$$

Operação Multiplicação ($\times$)

Na multiplicação e divisão as mantissas e características são tratadas separadamente.

$$\begin{array}{r} 31313142\ \underline{51} \\ \times 12315782\ \underline{65} \\ \hline \end{array}$$

mantissa $= 0.31313142 \times 0.12315782 = 0,038564583$

característica $= 51 + 65 - 50 = 66$, onde -50 é o desconto para compensar o ajuste $+50$ em cada ajuste do expoente da representação.

A produto é:

$$31313142\ \underline{51} \times 12315782\ \underline{65} = 038564583\ \underline{66}$$

Com a normalização obtém-se o resultado: $38564583\ \underline{65}$

Operação Divisão ($\div$)

Na divisão as mantissas e características são tratadas separadamente.

$$\frac{31313142\ \underline{51}}{12315782\ \underline{65}}$$

$$\text{mantissa} = \frac{0.31313142}{0.12315782} = 2.5425216198208$$

característica $= 51 - 65 + 50 = 36$, onde $+50$ é adicionado para compensar o cancelamento $+50$ do ajuste do expoente em cada representação.
A divisão é:

$$\frac{31313142\ \underline{51}}{12315782\ \underline{65}} = 2.5425216198208\ \underline{36}$$

Com a normalização obtém-se o resultado: $25425216\ \underline{37}$

1.3 Análise de erros

Resultados exatos dos cálculos são um supremo ideal em análise numérica. Quatro tipos de erro afetam a exatidão dos cálculos [**Ueberhuber1997**]: erros de modelo, erros de dados, erros de algoritmos e erros de arredondamento. A maioria da literatura em língua inglesa faz classificação diferente. Estes erros não são conseqüências de equívocos ou decisões precipitadas. Diferente, por exemplo, de erros de programação, eles são inevitáveis. Em muitos casos eles podem ser antecipados, e requerimentos de exatidão podem ser impostos, i.e., eles podem ser controlados para permanecerem abaixo de certos limites de erros. Os limites de erro são parte da especificação do problema numérico:

$$|\text{Erros do modelo + erros dos dados + erros de algoritmos + erro de arredondamento}| \leq \text{tolerância}$$

Todos os erros relevantes têm que ser identificados e seus efeitos nos resultados numéricos devem ser avaliados. Nas seções seguintes os quatro tipos de erros são caracterizados.

1.3.1 Erros de modelo

Em qualquer processo de modelagem, várias grandezas são desprezadas. O modelo resultante é uma abstração da realidade e vários modelos podem ser utilizados. O desvio inevitável entre o modelo e o objeto modelado é denotado por erro de modelagem. É necessário estimar a magnitude dos efeitos dos erros de modelagem para garantir os requisitos de tolerância de erro. Normalmente tais estimativas não são obtidas pois os fatores envolvidos são desconhecidos e não quantificados.

1.3.2 Erro de dados

Geralmente modelos não são para uma aplicação específica, mas para uma classe de aplicações similares. Uma instância é identificada por valores de parâmetros do modelo. Por exemplo, o comprimento l, o deslocamento angular inicial θ e a constante gravitacional g (que depende da localização geográfica) são parâmetros do modelo matemático do pêndulo simples. Devido a medições inexatas e outros fatores, os valores usados para parâmetros do modelo diferem do verdadeiro valor (normalmente desconhecido); isto é chamado de erro de dados. Os impactos dos erros de dados são objetos de análise.

1.3.3 Erro de algoritmo

Quando um problema matemático não pode ser resolvido analiticamente usando manipulações algébricas, então pode ser tentada uma solução por algoritmo numérico. No desenvolvimento de algoritmos numéricos são feitas simplificações antes que uma formulação finita do problema possa ser obtida para que o esforço computacional requerido seja reduzido a um nível razoável. O desvio resultante dos resultados obtidos pelo algoritmo dos da solução do problema matemático é denotado por erro de algoritmo.

Exemplo:

A solução do sistema de equações

$$Ax = b, \qquad A \in \mathbb{R}^{nxn}, \qquad b, x \in \mathbb{R}^n$$

requer um esforço computacional proporcional a n^3. Se uma solução aproximada $\widetilde{x}$ satisfazendo

$$||A\widetilde{x} - b|| < \epsilon$$

é suficiente, o custo computacional pode ser reduzido significativamente. Se A não possui estrutura especial, então somente

$$k \approx \sqrt{\kappa_2}\frac{ln(2/\epsilon)}{2}, \qquad \kappa_2 := \|A\|_2 \left\|A^{-1}\right\|_2$$

multiplicações matriz-vetor são necessárias para solução iterativa [**Traub1984**]. O número κ_2 é o número condição euclidiano da matriz A (ver seção 13.8 [**Ueberhuber1997**]).

1.3.4 Erro de truncamento

Algoritmos numéricos implantados em um computador podem somente realizar uma seqüência finita de operações aritméticas (adição, subtração, multiplicação, divisão e lógicas). Para calcular funções predefinidas sin, exp, ln, ... somente uma seqüência finita de operações aritméticas são executados pelo computador. O erro devido a troca de um processo infinito por uma seqüência finita de operações aritméticas é chamado de erro de truncamento.

Exemplo 1.18

A troca da série infinita da exponencial:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad com \quad P_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}$$

produz um erro de truncamento

$$e_{trunc}(x) := P_n(x) - cos(x)$$

que cresce com a distância entre x e zero.

1.3.5 Erros de discretização

O erro resultante de uma troca de informação contínua por informação discreta, num processo de amostragem, é referido como erro de discretização. Muitos autores estendem o termo erro de truncamento para incluir o erro de discretização.

Exemplo 1.19

O cálculo de uma integral definida

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

são aproximados por somas finitas envolvendo uma malha de pontos $x_i \in [a, b],\ i = 1, 2, \cdots n$ que pertencem ao conjunto de números de pontos flutuantes e as correspondentes avaliações aproximadas das funções $\{f(x_1), \cdots f(x_n)\}$. Os erros de arredondamento comprometem a exatidão dos resultados quando tenta-se minimizar os erros de truncamento refinando-se a malha de discretização.

1.3.6 Erros de arredondamento

Um computador fornece somente um conjunto finito de números: inteiros, e número de ponto-flutuante com mantissa de comprimento fixo. Desta forma as operações feitas num programa de computador não são geralmente executadas exatamente. Cada passo mapeia seu resultado em um dos números de ponto-flutuante disponíveis, normalmente o mias próximo. A diferença entre o resultado exato e o resultado arredondado de uma operação á chamado de erro de arredondamento. O efeito dos erros de arredondamento acumulados sobre o resultado final do método de aproximação é chamado de efeito de erro de arredondamento.

1.3.7 Erros em cálculos práticos

A fórmula do volume de uma esfera é: $V = \frac{4}{3} \cdot \pi \cdot r^3$ A expressão decimal $V = 1.333 \cdots \times 3.1415926535 \cdots \times r^3$ contém infinidades de algarismos, dos quais só utiliza-se alguns nos cálculos.

Exemplo: $r = 1.37m$ é um valor aproximado fornecido por uma medida, então

$V = 1.33 \times 3.14 \cdots \times (1.37)^3 \approx 10,74m^3$

A utilização de mais algarismos nos cáculos tem maior custo computacional e gera a ilusão de se ter obtido uma grande aproximação. Na realidade têm-se a geração de algarismos desprovidos de qualquer valor. Quando se trabalha com números aproximados é necessário avaliar a exatidão dos resultados.

1.3.8 Definições das Medidas dos Erros

As medidas de erro (ou de exatidão) mais utilizadas são o erro absoluto e o erro relativo.

Se x é um número real e $\tilde{x}$ sua aproximação, então:

O **erro absoluto** da aproximação $\tilde{x}$ é definido como

$$|x - \tilde{x}|$$

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

$$\frac{|x - \tilde{x}|}{|x|}, \text{se } x \neq 0$$

o erro relativo é adimensional e, muitas vezes, é expresso como percentual. O erro relativo em porcentagem da aproximação é dado por

$$\frac{|x - \tilde{x}|}{|x|} \times 100\%, \text{se } x \neq 0$$

1.3.9 Arredondamento e Truncamento

Durante as computações numéricas, após a normalização, os resultados dos cáculos são arredodados ou truncados para t dígitos, onde se obtém o resultado aproximado. Assumindo que todos os dígitos envolvidos no aproximação resultante são exatos, têm-se:

Erro de Truncamento $< \beta^{-t}$

Erro de Arredondamento $< \frac{1}{2} \times \beta^{-t}$

onde $\beta = 10$ para base padrão decimal.

No trucamento desprezado os dígitos após $t-$*dígitos. O arredodamento simétrico é o considerado.*

1.4 Propagação de erros

Nesta seção vemos estudar como o erro de computações numeércas se propagam (arredondamentos/truncamentos).

1.4.1 Propagação de Erros nas Operações Aritméticas: Adição

Suponha que um resultado x é resultante da adição de duas quantidades a e b

$$x = a + b$$

Considere Δa e Δb serem os erros absolutos na mediada de a e b e Δx ser o valor absoluto do erro em x.

$$x \pm \Delta x = (a \pm \Delta a) + (b \pm \Delta b) = (a + b) \pm (\Delta a + \Delta b)$$

$$= x \pm (\Delta a + \Delta b)$$

Então

$$\Delta x = \Delta a + \Delta b$$

Portanto o erro máximo absoluto em $x =$ erro máximo asoluto em $a +$ erro máximo asoluto em b.

1.4.2 Propagação de Erros nas Operações Aritméticas: Subtração

Suponha que um resultado x é resultante da subtração de duas quantidades a e b

$$x = a - b$$

Considere Δa e Δb serem os erros absolutos na mediada de a e b e Δx ser o valor absoluto do erro em x.

$$x \pm \Delta x = (a \pm \Delta a) - (b \pm \Delta b) = (a - b) \pm (\Delta a + \Delta b)$$

$$= x \pm (\Delta a + \Delta b)$$

Então

$$\Delta x = \Delta a + \Delta b$$

Portanto o erro máximo absoluto em $x =$ erro máximo asoluto em $a +$ erro máximo asoluto em b.

1.4.3 Propagação de Erros nas Operações Aritméticas: Multiplicação

Suponha que um resultado x é resultante da multiplicação de duas quantidades a e b

$$x = a \times b$$

Considere Δa e Δb serem os erros absolutos na medida de a e b e Δx ser o valor absoluto do erro em x.

$$\begin{aligned} x \pm \Delta x &= (a \pm \Delta a) \times (b \pm \Delta b) \\ &= (a \times b) \pm (a \cdot \Delta b + b \cdot \Delta a + \Delta a \cdot \Delta b) \\ &= x \pm (a \cdot \Delta b + b \cdot \Delta a + \Delta a \cdot \Delta b) \end{aligned}$$

Então

$$\frac{\Delta x}{x} = \frac{\cdot \Delta a}{a} + \frac{\Delta b}{b} + \frac{\Delta a}{a} \times \frac{\Delta b}{b} \approx \frac{\Delta a}{a} + \frac{\Delta b}{b}$$

Portanto o erro máximo relativo $\Delta x/x =$ erro máximo relativo $\Delta a/a +$ erro máximo ralativo $\Delta b/b$. Os produtos dos erros relativos $\Delta a/a \times \Delta b/b$ são considerados muito pequenos e negligenciados.

1.4.4 Propagação de Erros nas Operações Aritméticas: Divisão

Suponha que um resultado x é resultante da divisão de duas quantidades a e b

$$x = \frac{a}{b}$$

Considere Δa e Δb serem os erros absolutos na medida de a e b e Δx ser o valor absoluto do erro em x.

$$\begin{aligned} x \pm \Delta x &= \frac{(a \pm \Delta a)}{(b \pm \Delta b)} \\ &= \frac{a}{b} \cdot \frac{(1 \pm \frac{\Delta a}{a})}{(1 \pm \frac{\Delta b}{b})} \end{aligned}$$

Expandindo binomialmente

$$\begin{aligned} \left(1 \pm \frac{\Delta b}{b}\right)^{-1} &= \Bigg(1 \mp \frac{\Delta b}{b} \pm \text{termos contendo} \\ &\qquad \text{potências mais altas de } \frac{\Delta b}{b}\Bigg) \end{aligned}$$

1.4.5 Propagação de Erros: Generalização

Em muitos casos nós estimamos o erro de uma função $f(x_1, x_2, \dots, x_n)$ com erros individuais nas variáveis $(x_1, x_2, ..., x_n)$ conhecidos. Nós encontramos diretamente que

$$\Delta f = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n \tag{1.6}$$

onde os termos de ordem superior foram desprezados. O erro máximo é dado por

$$|\Delta f| = \left|\frac{\partial f}{\partial x_1}\right||\Delta x_1| + \left|\frac{\partial f}{\partial x_2}\right||\Delta x_2| + \cdots + \left|\frac{\partial f}{\partial x_n}\right||\Delta x_n| \tag{1.7}$$

O limite superior do erro é geralmente bastante pessimista, em computações práticas os erros têm uma tendência a cancelar [**Froberg65**]. Por exemplo, se 20000 números são arredondados com quatro casa decimais e adicionados, o erro máximo é

$$\frac{1}{2} \times 10^{-4} \times 20000 = 1$$

Um experimento com simulação Monte Carlo- por exemplo, gerando erros uniformemente distribuídos na quarta casa decimal, demonstra que a previsão acima é pessimista para o cenário examinado.

Exemplo 1.20

O volume de uma caixa é calculado pelo Comprimento, Largura e Altura, $V = C \times L \times A$. A caixa tem medidas de 10.5 ± 0.3m por 6.3 ± 0.2m por 4.8 ± 0.1m. Encontre o volume e a incerteza no volume.

V =(10.5m) · (6.3m) · (4.8m) = 317.52 m^3

$\Delta V = yz\Delta x + xz\Delta y + xy\Delta z$

$|\Delta V| = |yz| \cdot |\Delta x| + |xz| \cdot |\Delta y| + |xy| \cdot |\Delta z|$

$|\Delta V| = 6.2m \cdot 4.8m \cdot 0.3m + 10.5m \cdot 4.8m \cdot 0.2m + 10.5m \cdot 6.2m \cdot 0.1m = 25.518m^3$

Então V = (312.52 ± 25.52) m^3

Exemplo 1.21

Considere $L = x \cdot cos(\theta)$ para $x = (3.0 \pm 0.2)$cm, $\theta = (50 \pm 2)^\circ = (0.87266 \pm 0.034907)$ rad. Encontre L e sua incerteza. (obs: a incerteza no ancgulo deve ser em radianos !)

L= 3.0 cm cos 50° = 1.928 cm

$\Delta L = cos(\theta)\Delta x - x \cdot sen(\theta)\Delta\theta$

Tomando o valor absuluto de cada termo

$|\Delta L| = |cos(\theta)| \cdot |\Delta x| + |x \cdot sen(\theta)| \cdot \Delta\theta$

$= cos(50^\circ)$·(0.2cm) + (3cm)·$sen(50^\circ) \cdot 0.034907 = 0.20878$cm

Então L = (1.93 ± 0.21) cm

1.4.6 Cancelamento numérico

Devido ao comprimento limitado das palavras em computadores, e em conseqüência do uso da aritmética do ponto-flutuante com representação normalizada, as leis da aritmética elementar não são satisfeitas. Os efeitos do uso da aritmética do ponto-flutuante serão vistos em alguns exemplos que seguem.

Os exemplos a seguir violam a lei associativa da adição:

Exemplo 1.22

(usando-se uma máquina com quatro dígitos decimais na mantissa da representação)

$x = 9.909 \quad y = 1.000 \quad z = -0.990$

$(x + y) + z = 10.90 + (-0.990) = 9.910$

$x + (y + z) = 9.909 + (0.010) = 9.919$

Exemplo 1.23

(usando-se uma máquina com quatro dígitos decimais na mantissa da representação)

$x = 4561 \quad y = 0.3472$

$(y + x) - x = (0.3472 + 4561) - 4561 = 4561 - 4561 = 0.0000$

$y + (x - x) = 0.3472 + (4561 - 4561) = 0.3472 + 0.0000 = 0.3472$

Vejamos agora um exemplo (usando-se uma máquina com quatro dígitos decimais na mantissa da representação) que viola a lei distributiva.

Exemplo 1.24

$x = 9909 \quad y = -1.000 \quad z = 0.999$

$(x \times y) + (x \times z) = -9909 + 9899 = -10.00$

$x \times (y + z) = 9909 \times (-0.0001) = -9,909$

Exemplo 1.25

A equação do segundo grau $x^2 - b \cdot x + \epsilon = 0$ tem duas soluções:

$$x_1 = \frac{b + \sqrt{b^2 - 4\epsilon}}{2} \text{ e } x_2 = \frac{b - \sqrt{b^2 - 4\epsilon}}{2}$$

Se $b > 0$ e $\epsilon << b$, x_2 é expresso como a diferença de dois números praticamente iguais e poderá perder muitos dígitos significativos.

Se x_2 for reescrito como:

$$x_2 = \frac{\epsilon}{x_1} = \frac{2\epsilon}{b + \sqrt{b^2 - 4\epsilon}}$$

a raiz é aproximadamente $\frac{\epsilon}{b}$ sem perda de dígitos significativos.

Usando-se uma máquina com quatro dígitos decimais na mantissa da representação:

$b = 300.0 \quad \text{e} \quad \epsilon = 1.000$
$\sqrt{90000 - 4.000} = 300.0$

$$x_1 = \frac{600.0}{2.000} = 300.0$$

$$x_2 = \frac{300.0 - 300.0}{2.000} = \frac{0.0000}{2.000} = 0.0000$$

usando a relação $x_2 = \frac{\epsilon}{x_1} = \frac{1.000}{300.0} = 0.0033$ é um resultado mais preciso.

Exemplo 1.26

Sabe-se que para x grande $senh(x) \cong cosh(x) \cong \frac{e^x}{2}$.
Se quisermos calcular e^{-x} podemos dizer que $e^{-x} = cosh(x) - senh(x)$, o que conduz a um cancelamento perigoso.
Por outro lado $e^{-x} = \frac{1}{cosh(x) + senh(x)}$ fornece resultados mais precisos.

$$senh(x) = \frac{e^x - e^{-x}}{2} \ \& \ cosh(x) = \frac{e^x + e^{-x}}{2}$$

Capítulo Execícios

1. Converta o número **1597,41** para base binária com oito dígitos após a vírgula (oito dígitos binários).

2. Do número binário encontrado na resposta da questão (1) utilizar tabelas de conversão para obtenção da sua representação octal e hexadecimal:

   $(1597{,}41)_{(2)} =$ _____ (8)

   $(1597{,}41)_{(2)} =$ _____ (16).

   Obs: $(1597, 41)_{(2)}$ *é o número binário obtido na questão (1).*

3. Converta os seguintes números octal (base 8) e hexadecimal (base 16) em binário usando tabelas de conversão correspondentes:

   $267, 45_{(8)} =$ _____ (2)

   $35\text{E}, 9\text{D}_{(16)} =$ _____ (2).

4. Converta para decimal os números seguintes

   $1202{,}2_{(3)} =$ _____ (10)

   $11010011{,}1011_{(2)} =$ _____ (10).

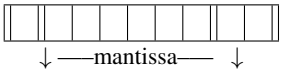
   $67{,}34_{(8)} =$ _____ (10)

   $3\text{C}{,}6\text{A}_{(16)} =$ _____ (10)

5. Escreva os números abaixo na notação normalizada $\mathbb{F}_N(10, 5, -50, 49)$.

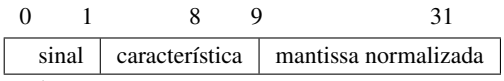
   a) -184.57
   b) 0.235
   c) -0.00379155445
   d) 0.0000471983
   e) 3194.74

6. Armazenar na memória de uma máquina decimal, utilizando a notação $\mathbb{F}_N(10, 6, -50, 49)$, o número $-753, 937486$ conforme esquema descrito abaixo:

   O sinal do número é codificado como $(+, -)$.

7. Represente o número -198.35 numa palavra de 32 *bits* conforme esquema representado abaixo.

   onde:

   sinal = sinal codificado 0 se $(+)$ e 1 se $(-)$;

   característica = 127 + expoente (convertido em binário);
   mantissa = fração binária normalizada.

8. Realize as operações aritmética abaixo na notação nomalizada $\mathbb{F}_N(10, 4, -50, 49)$:

a) 4154 $\underline{28}$ + 4314 $\underline{28}$ = _____,
b) 7342 $\underline{63}$ + 8919 $\underline{63}$ = _____,
c) 2157 $\underline{74}$ + 7581 $\underline{76}$ = _____,
d) 8821 $\underline{83}$ - 8756 $\underline{83}$ = _____.

9. Dado os números $A = 582143\ \underline{37}$ e $B = 216842\ \underline{86}$ armazenados na memória de uma máquina decimal utilizando a notação $\mathbb{F}_N(10, 6, -50, 49)$, realizar as seguintes operações:

a) $A \cdot B$ = _____,
b) $\dfrac{B}{A}$ = _____.

10. Assuma que os números 1.144 e 61.732 possuam cada um, um erro absoluto máximo de 0.0005. Qual o erro relativo máximo em cada número? Qual o erro absoluto máximo nas suas operações aritméticas (soma, subtração, multiplicação e divisão)?

11. Qual o erro absoluto máximo resultado de $f(x) = e^{-x}cos(x)$ ao se usar o valor de $x = -0.9 \pm 0.005$?

Exercícios Suplementares

1) Converta os seguintes números decimais para sua forma binária:

a) 35 b) 2345 c) 0.1218 d) 67.67 e) 95 f) 2500
g) 2000 h) 655 i) 722 j) 3.6×10^{21} l) 231 m) 2.5×10^{-18}

2) Converta os números binários para suas formas octal, hexadecimal e decimal:

a) 101101_2 b) -110101011_2 c) -0.1101_2
d) 0.111111101_2 e) 0.0000101_2 f) 10101_2
g) -11101011011_2 h) -0.1100001_2 i) 0.101100111101_2
j) 0.001100101_2

3) Reescreva os números seguintes na representação do ponto-flutuante normalizada:

a) 27.534 b) -89.901 c) 18×10^{21} d) 1.3756×10^{-7}
e) 11.0111_2 f) -111.0101_2 g) 0.00101_2 h) 111010101_2

4) Seja o número seguinte em ponto-flutuante num computador de 32 bits:

0010.0101.0000.0001.0001.1001.1100.1110

Se o primeiro bit é o sinal do número, os oito seguintes a característica obtida com adição de 128 ao expoente do número ponto-flutuante, e os 23 restantes são a mantissa, responda às questões seguintes:

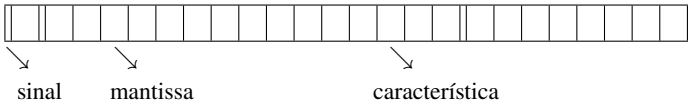
a) O número está normalizado? Se não o normalize.
b) Qual o sinal do número?
c) O valor absoluto do número é menor que 1?

5) Repita a questão 4 com o número:

1000.0000.0110.1101.1010.1101.1011.0110

6) Para a representação da questão 4, quais são aproximadamente o maior e o menor número, o menor número positivo e o próximo menor número positivo.

7) Represente os números binários da questão 2 na maquina binária que utiliza o seguinte esquema de representação de ponto-flutuante:



a) o bit de sinal é codificado 0 se o número é positivo e 1 se o número é negativo.
b) a característica é obtida com adição de 128 ao expoente do número ponto-flutuante.

8) Converter para base 10 os valores representados na máquina binária da questão 7) acima:

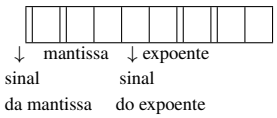
a)

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

b)

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

9) Seja um sistema de aritmética de ponto-flutuante na base decimal com quatro dígitos na mantissa e dois na característica, 1 digito de sinal da mantissa e 1 digito sinal da característica.



O sinal é codificado (+) se o número é positivo e (-) se o número é negativo.

Dados os números:

$x = 0.77237 \quad y = 0.2145 \times 10^{-3} \quad z = 0.2585 \times 10^{1}$

Efetue as seguintes operações:

a) $x + y + z$ b) $x - y - z$ c) x/y d) $(xy)/z$ e) $x(y/z)$

10) Use a aritmética do ponto-flutuante, com a representação da questão 9 acima, para somar e subtrair os seguintes pares de números:

a) 5.414234 e 2.27531 b) 5.414234 e 22.7531
c) 54.67 e 0.328 d) 5.4×10^{-8} e 3.14×10^{-5}

11) Use a aritmética do ponto-flutuante, com a representação da questão 9 acima, para realizar as operações aritméticas seguintes:

a) 3.14×7.47 b) 75.81×8.15 c) $1.35 \div 28.5$ d) $4000 \div 150$

12) Calcular as cotas dos erros absolutos e relativos que se comete ao se tomar como valores de:

a) 22/7 b) 333/116 c) 355/113 d) $\sqrt{3} + \sqrt{2}$

13) Ao se calcular $cos(x) \cong 1 - \frac{x^2}{2!} - \frac{x^4}{4!} - \frac{x^6}{6!}$ para $x = 5/7$, quais são os erros: inicial, de truncamento, de arredondamento e total cometidos quando se realiza os cálculos arredondados em duas casas decimais.

Capítulo - Resolução de Sistemas de Equações Lineares

Este capítulo dedica-se a solução de n equações algébricas lineares em n incógnitas. É muito comum a solução numérica de problemas práticos que envolvem equações simultâneas. Além disso, sistemas lineares decorrentes de problemas físicos são geralmente muito grandes, que consomem muitos recursos computacionais. Portanto há algoritmos desenvolvidos para a solução dos grandes conjuntos de equações, sendo cada um deles adaptado a um determinado forma dos coeficientes da matriz (esparsas, simétrica, em faixas, etc). Coleções bem conhecidas destas rotinas são: NAG, LAPACK, IMSL e NETLIB. Está fora do escopo deste livro discutir todos os algoritmos especiais. A opção é apresentar alguns dos métodos básicos de solução.

Um sistema de equações lineares, com m equações e n variáveis, é escrito geralmente como:

$$\left\{\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}\right. \tag{2.1}$$

onde
$a_{ij}, i = 1 \cdots m, j = 1 \cdots n$ são coeficientes constantes
$x_j, j = 1 \cdots n$ são varáveis
$b_i, i = 1 \cdots m$ são constantes

A resolução de um sistema linear determina os valores de $x_j, j = 1 \cdots n$ que satisfazem as m equações.

O sistema linear pode ser representado, usando a notação matricial, por:

$$A \cdot \mathbf{x} = \mathbf{b} \tag{2.2}$$

onde

$$A = \left(\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array}\right) \quad \textit{é a matriz dos coeficientes} \tag{2.3}$$

$$\mathbf{x} = \left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right) \quad \textit{é a matriz coluna das variáveis} \tag{2.4}$$

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \quad é\ a\ matriz\ coluna\ das\ constantes. \tag{2.5}$$

Denota-se por x^* o vetor solução e $\tilde{x}$ uma solução aproximada do sistema linear $A \cdot \mathbf{x} = \mathbf{b}$.

Uma representação útil das equações para fins de computações manuais (usando calculadoras, mas sem usar programação) é o da matriz de coeficiente aumentado, obtida pela adjacência do vetor constante b a matriz de coeficiente da seguinte forma:

$$A\,|\mathbf{b} = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right]$$

As situações que podem ocorrer com relação ao número de soluções de um sistema linear são:

1. Solução única (sistema determinado);
2. Infinitas soluções (sistema indeterminado);
3. Não admite soluções (sistema incompatível).

Será apresentado métodos numéricos para encontrar a solução única de sistemas lineares $n \times n$. Os métodos numéricos são divididos em dois grupos: métodos diretos e métodos iterativos.

- Métodos Diretos
  - **Gauss**
  - **Jordan**
  - **Decomposição LU (Doolittle, Crout e Choleski)**
- Métodos Iterativos
  - Estacionários
    - **Jacobi**
    - **Gauss-Seidel**
    - SOR
  - Não-estacionários
    - CG
    - (CGNE,CGNR,CGS)
    - BiCG
    - BiCGSTAB

SOR (Successive Overrelaxation) - Sobre-Relaxação sucessiva

CG (Conjugate Gradient) - Gradiente Conjugado

CGNR (Conjugate Gradient on the Normal equations to minimize the Residual) - Gradiente Conjugado sobre equações normais para minimizar o Resíduo

CGNE (Conjugate Gradient on the Normal equations to minimize the Error) - Gradiente Conjugado sobre equações normais para minimizar o Erro

CGS (Conjugate Gradient Squared) -Qudarado do Gradiente Conjugado

BiCG (Biconjugate gradient) - Gradiente Bi-Conjugado

BiCGSTAB (Bi-conjugate gradient stabilized method) - Gradiente Bi-Conjugado Estabilizado

Nota

O métodos destacados com negrito serão apresentados a seguir

Métodos diretos fornecem a solução exata x^*, não considerando erros de arredondamento, na solução do sistema linear, após um número determinado de operações.

Métodos iterativos geram uma sequência de vetor $\mathbf{x}^{(k)}$, a partir de uma aproximação inicial $\mathbf{x}^{(0)}$. Sob determinadas condições, esta sequência converge para a solução .

2.1 Métodos diretos

Todos os métodos de solução de sistemas lineares estudados nos 1º e 2º graus são diretos. A Regra de Cramer aplicada a resolução de um sistema $n \times n$ envolve o cálculo de $(n+1)$ determinantes de ordem n, então o número total de operações necessárias é cerca de $\approx e \times (n+1)!$ com $n \rightarrow \infty$ [**Suli2003**] [1], https://books.google.com.br/books?id=hj9weaqJTbQC utilizando o desenvolvimento por cofatores no cálculo dos determinantes.

Se n for igual a 30 serão efetuadas $31!e \approx 2.23510^{34}$ operações aritméticas usando o desenvolvimento por cofatores no cálculo dos determinantes. Assim um computador que efetuar um bilhão (10^9) de multiplicações por segundo levaria $\frac{2.235 \times 10^{34}}{10^9} = 2.235 \times 10^{25}$ segundos, ou seja 7.09×10^{17} anos para efetuar as multiplicações necessárias. De acordo com a teoria do *Big Bang* o Universo foi criado por uma violenta explosão cerca de $12.5 \pm 3 \times 10^9$ anos atrás. Métodos mais eficientes são necessários, pois problemas práticos exigem a resolução de sistemas lineares de grande porte.

A Tabela (2.1) lista três métodos diretos populares, todos usam operações elementares para produzir sua própria forma final de equações fáceis de resolver.

Operações Elementares sobre um Sistema de Equações Lineares:

a) Trocar a posição das equações;
b) Multiplicar uma equação por uma constante não nula;
c) Multiplicar uma equação por uma constante e adicionar a outra equação e, então, substituir esta nova equação por uma das existentes.

Tabela 2.1: Métodos diretos populares

Método	Forma Inicial	Forma final
Eliminação de Gauss	$A \cdot \mathbf{x} = \mathbf{b}$	$T \cdot \mathbf{x} = \mathbf{c}$
Eliminação de Gauss-Jordan	$A \cdot \mathbf{x} = \mathbf{b}$	$I \cdot \mathbf{x} = \mathbf{c}$
Decomposição LU	$A \cdot \mathbf{x} = \mathbf{b}$	$LU \cdot \mathbf{x} = \mathbf{b}$

2.1.1 Método da Eliminação de Gauss - Triangularização

O método da Eliminação de Gauss consiste em transformar o sistema de equações lineares original num sistema triangular superior equivalente que tem solução imediata, através do método da substituição retroativa, como vimos acima. Operações elementares produzem sistemas lineares equivalentes- que possuem a mesma solução do sistema original.

Descreveremos a seguir como o método de eliminação de Gauss usa as operações elementares para triangularizar um sistema de equações lineares. Para que isto ocorra é preciso supor que $detA \neq 0$, onde A é a matriz dos coeficientes.

Considerando que $detA \neq 0$ é sempre possível reescrever o sistema linear de forma que o elemento da posição a_{ii} seja diferente de zero, usando somente a operação elementar de troca de linha.

Seja a representação do sistema, com $a_{11} \neq 0$, pela matriz aumentada abaixo:

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ a_{21}^{(0)} & a_{21}^{(0)} & \cdots & a_{11}^{(0)} & b_2^{(0)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1}^{(0)} & a_{n2}^{(0)} & \cdots & a_{11}^{(0)} & b_n^{(0)} \end{array}\right]$$

[1] Süli, E. and Mayers, D.F. An Introduction to Numerical Analysis, Cambridge University Press, 2003

Vamos realizar a triangularização por etapas:

1ª ETAPA - Colocar zero abaixo do elemento da diagonal $a_{11}^{(0)}$. Ao final da 1ª etapa tem-se a matriz aumentada abaixo:

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ 0 & a_{21}^{(1)} & \cdots & a_{11}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{11}^{(1)} & b_n^{(1)} \end{array}\right]$$

Para isto subtraímos da i-ésima equação da 1ª equação multiplicada por $m_{i1} = \dfrac{a_{i1}^{(0)}}{a_{11}^{(0)}}, \quad i = 2 \cdots n$. Os m_{i1} são os multiplicadores e o elemento $a_{11}^{(0)}$ é chamado de pivô da primeira etapa. Sendo assim, $L_i = L_i - m_{i1}L_1 \ , \quad i = 2 \cdots n$, serão as novas linhas que substituirão as linhas no processo de eliminação da 1ª etapa.

2ª ETAPA - Colocar zero abaixo do elemento da diagonal $a_{22}^{(1)}$. Ao final da 2ª etapa tem-se a matriz aumentada abaixo:

$$\left[\begin{array}{ccccc|c} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & b_3^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

Para isto subtraímos da i-ésima equação da 2ª equação multiplicada por $m_{i2} = \dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3 \cdots n$. Os m_{i2} são os multiplicadores e o elemento $a_{22}^{(1)}$ é chamado de pivô da segunda etapa. Sendo assim, $L_i = L_i - m_{i2}L_2 \ , \quad i = 3 \cdots n$, serão as novas linhas que substituirão as linhas no processo de eliminação da 2ª etapa.

(n-1)ª ETAPA - Colocar zero abaixo do elemento da diagonal $a_{n-1,n-1}^{(n-2)}$ concluindo o processo de triangularização. Ao final da (n-1)ª etapa, tem-se a matriz aumentada abaixo:

$$\left[\begin{array}{ccccc|c} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & b_3^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]$$

Para isto subtraímos da n-ésima equação da (n-1)ª equação multiplicada por $m_{n,n-1} = \dfrac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}$. O $m_{n,n-1}$ é o multiplicador e o elemento $a_{n-1,n-1}^{(n-2)}$ é chamado de pivô da (n-1)ª etapa. Sendo assim, $L_n = L_n - m_{n,n-1}L_{n-1}$, é a nova linhas que substituirá a última linha no processo de eliminação da (n-1)ª etapa.

Agora o sistema é triangular superior e equivalente ao sistema de equações lineares original. Procede-se a substituição retroativa para resolução do sistema triangular, e então, obter-se a solução do sistema e completa-se o algoritmo.

Uma medida da eficiência de um algoritmo é o número de operações aritméticas necessárias para obter a solução [Conte1980] [2] https://books.google.com.br/books?id=tNBTDwAAQBAJ.

[2] Conte, S.D. and De Boor, C., Elementary Numerical Analysis: An Algorithmic Approach, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2017

A fase de eliminação efetua $\frac{n\cdot(n-1)}{2}$ divisões, $\frac{(n^3-n)}{3}$ multiplicações, e $\frac{(n^3-n)}{3}$ adições .

Para resolver o sistema triangular superior são efetuadas n divisões, $\frac{n\cdot(n-1)}{2}$ multiplicações, e $\frac{n\cdot(n-1)}{2}$ adições.

Então o total de operações para se resolver um sistema linear pelo método de Eliminação de Gauss é $\frac{(4n^3+9n^2-7n)}{6} = \frac{2}{3}n^3+\frac{3}{2}n^2-\frac{7}{6}n$. Assim um computador que efetuar um bilhão (10^9) de operações por segundo levaria $5.334\cdot 10^3$ segundos $\cong 1.48$ horas para resolver um sistema de equações lineares 20000x20000.

Exemplo 2.1

Resolver o sistema de equações lineares abaixo pelo método de eliminação de Gauss:

$$\begin{cases} 3x+2y+4z &= 1\\ x+\ y-2z &= 0\\ 4x+3y-2z &= 2 \end{cases}$$

Trabalharemos com a matriz de coeficientes ampliada com o vetor constante:

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

1ª ETAPA:

Pivô: $a_{11}^{(0)} = 3$

$m_{21} = 1/3$

$m_{31} = 4/3$

$L_2 \leftarrow L_2 - m_{21}\cdot L_1$

$L_3 \leftarrow L_3 - m_{31}\cdot L_1$

Assim tem-se após a 1ª ETAPA

$$\left[\begin{array}{ccc|c} 3 & 2 & 4 & 1\\ 0 & 1/3 & -10/3 & -1/3\\ 0 & 1/3 & -22/3 & 2/3 \end{array}\right]$$

2ª ETAPA:

Pivô: $a_{22}^{(1)} = 1/3$

$m_{32} = \dfrac{1/3}{1/3} = 1$

$L_3 \leftarrow L_3 - m_{32}\cdot L_2$

Assim tem-se após a 2ª ETAPA

$$\left[\begin{array}{ccc|c} 3 & 2 & 4 & 1\\ 0 & 1/3 & -10/3 & -1/3\\ 0 & 0 & -4 & 1 \end{array}\right]$$

Usando a notação para cálculos manuais tem-se mais resumidamente:

	3	2	4	1	**(1)**
(1/3)	1	1	−2	0	
(4/3)	4	3	−2	2	
	0	1/3	−10/3	−1/3	**(2)**
(1)	0	1/3	−22/3	2/3	
	0	0	−4	1	**(3)**

Os valores entre parêntesis à esquerda são os multiplicadores usados na eliminação. A matriz triangular é obtida usando-se a equação do pivô de cada etapa da eliminação que estão indicadas entre parêntesis à direita e em negrito.

Agora resolver $A \cdot \mathbf{x} = \mathbf{b}$ é equivalente a resolver $T \cdot \mathbf{x} = \mathbf{c}$

$$\begin{aligned} 3x + 2y + 4z &= 1 \\ 1/3y - 10/3z &= -1/3 \\ -4z &= 1 \end{aligned}$$

Logo:

$z = -1/4$
$y = -1 + 10z = -1 + 10 \cdot (-1/4) \Longrightarrow y = -14/4 = -7/2$
$3x = 1 - 2 \cdot (-7/2) - 4 \cdot (-1/4) = 1 + 7 + 1 = 9 \Longrightarrow x = 3$

Solução: $\{x = 3,\ y = -7/2,\ z = -1/4\}$

2.1.2 Método da Eliminação de Gauus-Jordan - Diagonalização

O método da Eliminação de Gauss-Jordan, ou simplesmente Jordan, consiste em transformar o sistema de equações lineares original num sistema diagonal equivalente que tem solução imediata. Ele é uma extensão do método de eliminação gaussiana. O método de eliminação de Jordan é usado para reduzir a matriz aumentada para a forma

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & 0 & \cdots & 0 & b_1^{(n)} \\ 0 & a_{21}^{(1)} & \cdots & 0 & b_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]$$

O método de eliminação de Jordan para diagonalizar um sistema de equações lineares será realizado de modo análogo ao da eliminação gaussiana.. Considerando que $detA \neq 0$ é sempre possível reescrever o sistema linear de forma que o elemento da posição a_{ii} seja diferente de zero, usando somente a operação elementar de troca de linha.

Seja a representação do sistema, com $a_{11} \neq 0$, dada pela matriz aumentada abaixo:

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ a_{21}^{(0)} & a_{21}^{(0)} & \cdots & a_{11}^{(0)} & b_2^{(0)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1}^{(0)} & a_{n2}^{(0)} & \cdots & a_{11}^{(0)} & b_n^{(0)} \end{array}\right]$$

Vamos realizar a diagonalização por etapas:

1ª ETAPA - Colocar zero abaixo do elemento da diagonal $a_{11}^{(0)}$. Ao final da 1ª etapa teremos a matriz aumentada abaixo:

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} & b_1^{(0)} \\ 0 & a_{21}^{(1)} & \cdots & a_{11}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{11}^{(1)} & b_n^{(1)} \end{array}\right]$$

Para isto subtraímos da i-ésima equação da 1ª equação multiplicada por $m_{i1} = \frac{a_{i1}^{(0)}}{a_{11}^{(0)}}, \quad (i = 1 \cdots n,\ i \neq 1)$. Os m_{i1} são os multiplicadores e o elemento $a_{11}^{(0)}$ é chamado de pivô da primeira etapa. Sendo assim, $L_i = L_i - m_{i1}L_1\ , \quad (i = 1 \cdots n,\ i \neq 1)$, serão as novas linhas que substituirão as linhas antes do processo de eliminação da 1ª etapa.

2ª ETAPA - Colocar zero acima e abaixo do elemento da diagonal . Ao final da 2ª etapa teremos a matriz aumentada abaixo:

$$\left[\begin{array}{ccccc|c} a_{11}^{(0)} & 0 & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

Para isto subtraímos da i-ésima equação da 2ª equação multiplicada por $m_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad (i = 1 \cdots n,\ i \neq 2)$. Os m_{i2} são os multiplicadores e o elemento $a_{22}^{(1)}$ é chamado de pivô da segunda etapa. Sendo assim, $L_i = L_i - m_{i2}L_2\ , \quad (i = 1 \cdots n,\ i \neq 2)$, serão as linhas que substituirão as linhas antes do processo de eliminação da 2ª etapa.

nª ETAPA - Colocar zero acima do elemento da diagonal concluindo o processo de diagonalização. Ao final da nª etapa, a última etapa, teremos a matriz aumentada abaixo:

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & 0 & \cdots & 0 & b_1^{(n-1)} \\ 0 & a_{21}^{(1)} & \cdots & 0 & b_2^{(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]$$

Agora o sistema é diagonal e equivalente ao sistema de equações lineares original.

A fase de eliminação efetua $n \cdot (n-1)$ divisões, $\frac{(n^3 - n)}{2}$ multiplicações, e $\frac{(n^3 - n)}{2}$ adições .

Para resolver o sistema diagonal são efetuadas n divisões.

Então o total de operações para se resolver um sistema linear pelo método de Jordan é $n^3 + n^2 - n$. O que assintoticamente é 50% superior ao Método de Gauss. Assim um computador que efetuar um bilhão (10^9) de operações por segundo levaria $8.000 \cdot 10^3$ segundos $\cong 2.22$ horas para resolver um sistema de equações lineares 20000x20000.

Exemplo 2.2

Resolver o sistema de equações lineares abaixo pelo método de eliminação de Jordan:

$$\begin{cases} 3x + 2y + 4z &= 1 \\ x + \ y - 2z &= 0 \\ 4x + 3y - 2z &= 2 \end{cases}$$

Trabalharemos com a matriz de coeficientes ampliada com o vetor constante:

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

1ª ETAPA:

Pivô: $a_{11}^{(0)} = 3$

$m_{21} = 1/3$

$m_{31} = 4/3$

$L_2 \leftarrow L_2 - m_{21} \cdot L_1$

$L_3 \leftarrow L_3 - m_{31} \cdot L_1$

Assim tem-se após a 1ª ETAPA

$$\left[\begin{array}{ccc|c} 3 & 2 & 4 & 1 \\ 0 & 1/3 & -10/3 & -1/3 \\ 0 & 1/3 & -22/3 & 2/3 \end{array}\right]$$

2ª ETAPA:

Pivô: $a_{22}^{(1)} = 1/3$

$m_{12} = \dfrac{2}{1/3} = 6$

$L_1 \leftarrow L_1 - m_{12} \cdot L_2$

$m_{32} = \frac{1/3}{1/3} = 1$

$L_3 \leftarrow L_3 - m_{32} \cdot L_2$

Assim tem-se após a 2ª ETAPA

$$\left[\begin{array}{ccc|c} 3 & 0 & 24 & 3 \\ 0 & 1/3 & -10/3 & -1/3 \\ 0 & 0 & -4 & 1 \end{array}\right]$$

3ª ETAPA:

Pivô: $a_{33}^{(2)} = -4$

$m_{13} = \dfrac{24}{-4} = -6$

$L_1 \leftarrow L_1 - m_{13} \cdot L_3$

$m_{23} = \dfrac{-10/3}{-4} = \dfrac{10}{12} = \dfrac{5}{6}$

$L_2 \leftarrow L_2 - m_{23} \cdot L_3$

Assim tem-se após a 3ª ETAPA

$$\left[\begin{array}{ccc|c} 3 & 0 & 0 & 9 \\ 0 & 1/3 & 0 & -7/6 \\ 0 & 0 & -4 & 1 \end{array}\right]$$

Usando a notação para cálculos manuais tem-se mais resumidamente:

$$
\begin{array}{rcccc}
 & 3 & 2 & 4 & 1 \\
-(1/3) & 1 & 1 & -2 & 0 \\
-(4/3) & 4 & 3 & -2 & 2 \\
\hline
-\left(\frac{2}{1/3}\right) & 3 & 2 & 4 & 1 \\
 & 0 & 1/3 & -10/3 & -1/3 \\
-\left(\frac{1/3}{1/3}\right) & 0 & 1/3 & -22/3 & 2/3 \\
\hline
-\frac{24}{-4} = (6) & 3 & 0 & 24 & 3 \\
-\frac{-10/3}{-4} = -\left(\frac{5}{6}\right) & 0 & 1/3 & -10/3 & -1/3 \\
 & 0 & 0 & -4 & 1 \\
\hline
 & 3 & 0 & 0 & 9 \\
 & 0 & 1/3 & 0 & -7/6 \\
 & 0 & 0 & -4 & 1
\end{array}
$$

Os valores entre parêntesis à esquerda são os multiplicadores (com sinal menos) usados na eliminação.

Agora resolver $A \cdot \mathbf{x} = \mathbf{b}$ é equivalente a resolver $A^{(2)} \cdot \mathbf{x} = \mathbf{b}^{(2)}$:

$$
\begin{array}{rcr}
3\text{x} & = & 9 \\
1/3\text{y} & = & -\ 7/6 \\
-\ 4\text{z} & = & 1
\end{array}
$$

Logo:

$x = 3$
$y = -7/2$
$z = -1/4$

Solução: $\{x = 3,\ y = -7/2,\ z = -1/4\}$

2.1.3 Métodos de decomposição LU

De forma semelhante a escalares. matrizes podem ser fatoradas em um produto de duas matrizes de infinitas maneiras. Portanto

$$A = B \cdot C \tag{2.6}$$

Quando $B = L$ e $C = U$ são matrizes triangular inferior L e superior U, respectivamente, a equação (2.6) resulta em

$$A = L \cdot U \tag{2.7}$$

As matrizes L e U são definidas por propriedades de seus elementos

$$L = \begin{cases} l_{ij}, & se\ i \geq j \\ 0, & se\ i < j \end{cases} \tag{2.8}$$

$$U = \begin{cases} u_{ij}, & se\ i \leq j \\ 0, & se\ i > j \end{cases} \tag{2.9}$$

Ao se especificar os elementos da diagonal de L ou de U a fatoração torna-se única. O procedimento baseado sobre elementos unitários na diagonal principal de L é chamado de *método de Doolittle*. O procedimento baseado sobre elementos unitários na diagonal principal de U é chamado de *método de Crout*.

Usa-se uma decomposição tal como (2.7) para resolver o sistema linear

$$A \cdot \mathbf{x} = (L \cdot U) \cdot \mathbf{x} = L \cdot (U \cdot \mathbf{x}) = \mathbf{b} \tag{2.10}$$

através da solução para o vetor y tal que

$$L \cdot \mathbf{y} = \mathbf{b} \tag{2.11}$$

e em seguida solucionando

$$U \cdot \mathbf{x} = \mathbf{y} \tag{2.12}$$

A vantagem de quebrar a solução de um sistema linear em dois é que a solução de um sistema linear é muito simples. A equação (2.11) pode ser resolvida por substituição direta, enquanto a equação (2.12) pode ser resolvida por substituição retroativa.

Um vez encontrada a decomposição LU de A, o sistema linear pode ser resolvido para muitos vetores independentes **b** do lado direito da equação com o cuidado de fazer um a cada vez. Esta é uma característica vantajosa diferenciada em relação aos métodos de eliminação de Gauss e Jordan. A decomposição LU pode ser usado para determinar a matriz inversa e determinantes.

2.1.3.1 Método de decomposição de Doolittle

No método LU de Doolittle , a matriz U é a matriz triangular superior obtida pela eliminação de Gauss. A matriz L é a matriz triangular inferior formada pelos multiplicadores da eliminação, obtido no processo de eliminação de Gauss, como elementos abaixo da diagonal, com elementos unidades sobre a diagonal principal.

Exemplo 2.3

Resolver o sistema de equações lineares abaixo pelo método de Doolittle:

$$\begin{cases} 3x + 2y + 4z &= 1 \\ x + y - 2z &= 0 \\ 4x + 3y - 2z &= 2 \end{cases}$$

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

1. usar 02 casas decimais e no mínimo com 02 dígitos significativos;
2. exibir cálculos detalhados das etapas de eliminação e substituição direta e retroativa.

Usando a notação para cálculos manuais agilizados:

$$\begin{array}{ccc} \textcircled{3} & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{array} \quad \textbf{(1)}$$

O $\textcircled{3}$ é o pivô da linha 1 na 1ª ETAPA. Então vamos eliminar tudo que estiver abaixo do pivô:

	③	2	4	**(1)**
$-(\frac{1}{3}) = -0.33$	1	1	-2	
$-(\frac{4}{3}) = -1.33$	4	3	-2	

Cálculos:

$$-0.33 \cdot (2, 4, 1) + (1, -2, 0) = (0.34, -3.32, -0.33)$$
$$-1.33 \cdot (2, 4, 1) + (3, -2, 2) = (0.34, -7.32, 0.67)$$

Assim tem-se após a 1ª ETAPA

	③	2	4	**(1)**
$-(\frac{1}{3}) = -0.33$	1	1	-2	
$-(\frac{4}{3}) = -1.33$	4	3	-2	
	3	2	4	
	$\boxed{0.33}$	(0.34)	-3.32	**(2)**
$-(\frac{0.34}{0.34}) = -1$	$\boxed{1.33}$	0.34	-7.32	

Destacando valores de $\boxed{l_{i,j}}$ obtidos utilizando o fatores utilizados para escalonamento-multiplicanbdo po -1.

O (0.34) é o pivô da linha 2 na 2ª ETAPA. Então vamos eliminar tudo que estiver abaixo do pivô:

Cálculos:

$$-1 \cdot (-3.32, -0.33) + (-7.32, 0.67) = (-4, 1)$$

Assim tem-se após a 2ª ETAPA

	3	2	4	**(1)**
$-(\frac{1}{3}) = -0.33$	1	1	-2	
$-(\frac{4}{3}) = -1.33$	4	3	-2	
	3	2	4	
	$\boxed{0.33}$	(0.34)	-3.32	**(2)**
$-(\frac{0.34}{0.34}) = -1$	$\boxed{1.33}$	0.34	-7.32	
	3	2	4	
	$\boxed{0.33}$	(0.34)	-3.32	
	$\boxed{1.33}$	$\boxed{1}$	(-4)	**(3)**

O (-4) é o pivô da linha 1 na 3ª ETAPA. Finalizada as etapas do escalonamento. Não temos nada eliminar abaixo do pivô.

Os fatores L e U são:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.33 & 1 & 0 \\ 1.33 & 1 & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} 3 & 2 & 4 \\ 0 & 0.34 & -3.32 \\ 0 & 0 & -4 \end{pmatrix}$$

$$LU \approx A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

Agora resolver $A\mathbf{x} = \mathbf{b}$ é equivalente a resolver $L(U\mathbf{x}) = \mathbf{b}$.

De $L\mathbf{y} = \mathbf{b}$:

$$\begin{cases} y_1 = 1 \\ 0.33y_1 + y_2 = 0 \\ 1.33y_1 + y_2 + y_3 = 1 \end{cases}$$

Logo:

$y_2 = -0.33 \cdot 1 = -0.33$
$y_3 = 2 - 1.33 \cdot 1 - (-0.33) = 1$
De $U\mathbf{x} = \mathbf{y}$:

$$\begin{cases} 3x_1 + 2x_2 + 4x_3 = 1 \\ 0.34x_2 - 3.32x_3 = -0.33 \\ -4x_3 = 1 \end{cases}$$

Logo:

$x_3 = -0.25$
$x_2 = \dfrac{-0.33 + 3.32 \cdot (-0.25)}{0.34} = -3.41$
$x_1 = \dfrac{1 - 2 \cdot (-3.41) - 4 \cdot (-0.25)}{3} = 2.94$

Solução: $\{x = 2.94,\ y = -3.41,\ z = -0.25\}$
Solução Exata: $\{x = 3,\ y = -7/2 = -3.5,\ z = -1/4 = -0.25\}$

2.1.3.2 Método de decomposição de Crout

Na prática a eliminação de Gauss é frequentemente realizada de acordo com um modificação sugerida por Crout. A partir do sistema (2.1) elimina-se da forma habitual para obter-se o seguinte sistema (presume-se que foram necessárias não permutações):

$$\begin{cases} x_1 + a'_{12}x_2 + a'_{13}x_3 + \cdots + a'_{1n}x_n = w_1 \\ \quad\ x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n = w_2 \\ \qquad\qquad\qquad\qquad\qquad\ \vdots \\ \qquad\qquad\qquad\qquad\ x_n = w_n \end{cases} \tag{2.13}$$

Uma determinada equação (i) neste sistema é o resultado da subtração da equação correspondente (2.1) de múltiplos das $(i-1)$ primeiras equações no sistema (2.13). Portanto a equação (i) em (2.1) pode ser escrita como uma combinação

linear das primeiras i equações em (2.13). A primeira equação de (2.1) é obtido pela multiplicação da equação correspondente (2.13) por um certo uma constante a'_{11}:

$$a'_{11}x_1 + a'_{11}a'_{12}x_2 + a'_{11}a'_{13}x_3 + \cdots + a'_{11}a'_{1n}x_n = a'_{11}w_1 \tag{2.14}$$

A segunda equação (2.1) é obtida através da multiplicação da primeira com a segunda equação em (2.13) por a'_{21} e a'_{22}, respectivamente, e adicionando-se os resultados:

$$\begin{array}{c} a'_{21}x_1 + (a'_{21}a'_{12} + a'_{22})x_2 + (a'_{21}a'_{13} + a'_{22}a'_{23})x_3 + \cdots \\ +(a'_{21}a'_{1n} + a'_{22}a'_{2n})x_n = a'_{21}w_1 + a'_{22}w_2 \end{array} \tag{2.15}$$

As equações restantes são formados de forma semelhante e, em seguida, os coeficientes são identificadas, e todas as constantes a'_{ij}, podem ser determinadas.

Particularmente, tem-se as seguintes equações para as do lado direito:

$$\left\{\begin{array}{lllllllll} a'_{11}w_1 & & & & & & & = & b_1 \\ a'_{21}w_1 & + & a'_{22}w_2 & & & & & = & b_2 \\ \vdots & & & & & & & & \\ a'_{n1}w_1 & + & a'_{n2}w_2 & + & \cdots & + & a'_{nn}w_n & = & b_n \end{array}\right. \tag{2.16}$$

Será introduzida as notações L e U da decomposição de Crout para o matrizes dos coeficientes de (2.16) e (2.13), respectivamente, incluindo os elementos unitários da diagonal de U. Então

$$l_{ij} = \left\{\begin{array}{l} a'_{ij} \ \ para \ \ i \geq j, \\ 0 \ \ para \ \ i < j, \end{array}\right. \quad u_{ij} = \left\{\begin{array}{l} 0 \ \ para \ \ i > j, \\ 1 \ \ para \ \ i = j, \\ a'_{ij} \ \ para \ \ i < j, \end{array}\right.$$

ou $L + (U - I) = A'$. Os sistemas (2.1), (2.13) e (2.16) são escritos como:

$$\left\{\begin{array}{l} A \cdot \mathbf{x} = \mathbf{b} \\ L \cdot \mathbf{w} = \mathbf{b} \\ U \cdot \mathbf{x} = \mathbf{w} \end{array}\right. \tag{2.17}$$

A terceira equação multiplicada com L e dá

$$L \cdot U \cdot \mathbf{x} = L \cdot \mathbf{w} = \mathbf{b} = A \cdot \mathbf{x}$$

Exemplo 2.4

Resolver o sistema de equações lineares abaixo pelo método de Crout:

$$\left\{\begin{array}{rcl} 3x + 2y + 4z & = & 1 \\ x + y - 2z & = & 0 \\ 4x + 3y - 2z & = & 2 \end{array}\right.$$

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

1. usar 02 casas decimais e no mínimo com 02 dígitos significativos;
2. exibir cálculos detalhados das etapas de eliminação e substituição direta e retroativa.

Usando a notação para cálculos manuais agilizados:

③	2	4	**(1)**
1	1	−2	
4	3	−2	

1ª ETAPA: O ③ da linha 1 é utilizado na uniformização dividindo todos os valores. Então vamos eliminar tudo que estiver abaixo do pivô:

1ª linha uniformizada:

	①	0.67	1.33	**(1)**
−1	1	1	−2	
−4	4	3	−2	

Gauss com uniformização:

1	0.67	1.33 **(1)**
1	0.33	−3.33
4	0.32	−7.32

Cálculos:

$\frac{1}{3} = 0.33 \cdot (3, 2, 4) = (1, 0.67, 1.33)$ uniformização

$-1 \cdot (0.67, 1.33) + (1, 1, -2) = (0.33, -3.33)$ nova linha 2

$-4 \cdot (0.67, 1.33) + (4, 3, -2) = (0.32, -7.32)$ nova linha 3

2ª ETAPA: O (0.33) da linha 2 é utilizado na uniformização dividindo todos os valores. Então vamos eliminar tudo que estiver abaixo do pivô:

2ª linha uniformizada:

	1	0.67	1.33	
	1	①	−10.09	**(2)**
−0.32	4	0.32	−7.32	

Gauss com uniformização:

1	0.67	1.33	
1	1	−10.09	**(2)**
4	0.32	−4.09	

Cálculos:

$\frac{1}{0.33} = 0.32 \cdot (0.33, -3.33) = (1, -10.09)$ uniformização

$-0.32 \cdot (-10.09) + (-7.32) = -4.09$ nova linha 3

3ª ETAPA: O (-4.09) da linha 2 é utilizado na uniformização dividindo todos os valores. Então, não tendo nada a eliminar abaixo do pivô, terminou a fatoração de Crout

3ª linha uniformizada:

	1	0.67	1.33	
	1	1	−10.09	
−0.32	4	0.32	①	**(3)**

Fim da Fatoração de Crout

Cálculos:

$\frac{-4.09}{-4.09} = 1$ uniformização

$$L = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 0.33 & 0 \\ 4 & 0.32 & -4.09 \end{pmatrix}$$

Nota *Os valores da diagonal de L são os utilizados na uniformização.*

$$U = \begin{pmatrix} 1 & 0.67 & 1.33 \\ 0 & 1 & -10.09 \\ 0 & 0 & 1 \end{pmatrix}$$

$$LU \approx A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

Agora resolver $A\mathbf{x} = \mathbf{b}$ é equivalente a resolver $L(U\mathbf{x}) = \mathbf{b}$.

De $L\mathbf{y} = \mathbf{b}$:

$$\begin{cases} 3y_1 = 1 \therefore y_1 = 0.33 \\ y_1 + 0.33y_2 = 0 \therefore y_2 = -\frac{0.33}{0.33} = -1 \\ 4y_1 + 0.32y_2 - 4.09y_3 = 2 \therefore y_3 = \dfrac{2 - 4 \times 0.33 - 0.32 \times (-1)}{-4.09} = -0.24 \end{cases}$$

De $U\mathbf{x} = \mathbf{y}$:

$$\begin{cases} x_1 + 0.67x_2 + 1.33x_3 = 0.33 \\ x_2 - 10.09x_3 = -1 \\ x_3 = -0.24 \end{cases}$$

Logo:

$$x_2 = -1 + 10.09 \times (-0.24) = -3.42$$

$$x_1 = 0.33 - 0.67 \times (-3.42) - 1.33 \times (-0.24) = 2.94$$

Solução: $\{x = 2.94,\ y = -3.42,\ z = -0.24\}$
Solução Exata: $\{x = 3,\ y = -7/2 = -3.5,\ z = -1/4 = -0.25\}$

Exemplo 2.5

Método de Crout ”Gauss com uniformização”: Resolução Ninja sem Exibir Cálculos

Uniformização $= \frac{1}{3}$	(3)	2	4	1
	1	1	−2	0
	4	3	−2	2
	1	0.67	1.33	0.33
−1	1	1	−2	0
−4	4	3	−2	2
	1	0.67	1.33	0.33
Uniformização $= \frac{1}{0.33}$	[1]	(0.33)	−3.33	−0.33
	[4]	0.32	−7.32	0.68
	1	0.67	1.33	0.33
	[1]	1	−10.09	−1
−0.32	[4]	[0.32]	−7.32	0.68
	1	0.67	1.33	0.33
	[1]	1	−10.09	−1
Uniformização $= \frac{1}{-4.09}$	[4]	[0.32]	(-4.09)	1
	1	0.67	1.33	0.33
	[1]	1	−10.09	−1
	[4]	[0.32]	1	−0.24

Logo:

$$x_3 = -0.24$$
$$x_2 = -1 + 10.09 \cdot (-0.24) = -3.42$$
$$x_1 = 0.33 - 0.67 \cdot (-3.42) - 1.33 \cdot (-0.24) = 2.94$$

Relação entre os Métodos de Doolittle e Crout

A transposta dos fatores L e U através da decomposição da transposta de A por um método de decomposição produz os fatores do outro método:

$$\begin{array}{l} \textit{se } L \cdot U = A^T \textit{ pelo Método de Doolittle} \\ \textit{então} \\ \qquad \begin{cases} L_{Crout} = U^T \\ U_{Crout} = L^T \end{cases} \end{array}$$

Exemplo 2.6

Encontre a decomposição de Crout através da decomposição de Doolittle da matriz

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

Partindo da matriz

$$A^T = \begin{pmatrix} 3 & 1 & 4 \\ 2 & 1 & 3 \\ 4 & -2 & -2 \end{pmatrix}$$

Realiza-se a eliminação da 1ª coluna através dos multiplicadores com sinal $(-)$ exibidos a frente:

$$\begin{matrix} \\ -2/3 \\ -4/3 \end{matrix} \begin{pmatrix} 3 & 1 & 4 \\ 2 & 1 & 3 \\ 4 & -2 & -2 \end{pmatrix}$$

Realiza-se a eliminação da 2ª coluna através do multiplicador com sinal $(-)$ exibido a frente:

$$\begin{matrix} \\ \\ 10 \end{matrix} \begin{pmatrix} 3 & 1 & 4 \\ 0 & 1/3 & 1/3 \\ 0 & -10/3 & -22/3 \end{pmatrix}$$

Obtendo-se

$$\begin{pmatrix} 3 & 1 & 4 \\ 0 & 1/3 & 1/3 \\ 0 & 0 & -4 \end{pmatrix}$$

Restaurando-se os elementos multiplicadores (para preservar armazenamento) tem-se

$$\begin{pmatrix} 3 & 1 & 4 \\ 2/3 & 1/3 & 1/3 \\ 4/3 & -10 & -4 \end{pmatrix}$$

A relação $LU = A^T$ tem a forma seguinte

$$\begin{pmatrix} 1 & 0 & 0 \\ 2/3 & 1 & 0 \\ 4/3 & -10 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 & 4 \\ 0 & 1/3 & 1/3 \\ 0 & 0 & -4 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 4 \\ 2 & 1 & 3 \\ 4 & -2 & -2 \end{pmatrix}$$

Logo:

$$L_{Crout} = U^T = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 1/3 & 0 \\ 4 & 1/3 & -4 \end{pmatrix}$$

e

$$U_{Crout} = L^T = \begin{pmatrix} 1 & 2/3 & 4/3 \\ 0 & 1 & -10 \\ 0 & 0 & 1 \end{pmatrix}$$

A relação $L_{Crout} \cdot U_{Crout} = A$ tem a forma seguinte

$$\begin{pmatrix} 3 & 0 & 0 \\ 1 & 1/3 & 0 \\ 4 & 1/3 & -4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2/3 & 4/3 \\ 0 & 1 & -10 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix}$$

2.1.3.3 Método de Decomposição de Choleski

No caso em que a matriz do sistema linear é simétrica podemos simplificar os cálculos da decomposição LU significativamente, levando em conta a simetria. O método de Cholesky é uma possibilidade, se baseia no seguinte teorema

Teorema 2.1

Se A é uma matriz simétrica positiva definida, então existe uma única matriz triangular inferior G com diagonal estritamente positiva, tal que

$$G \cdot G^T$$

♡

2.1.3.4 Esquema Prático para a Decomposição GG^T

Para obtermos a matriz G aplicamos a definição de produto e igualdade de matrizes. Seja então:

$$G \cdot G^T = \begin{pmatrix} g_{11} & 0 & 0 & 0 & \cdots & 0 \\ g_{21} & g_{22} & 0 & 0 & \cdots & 0 \\ g_{31} & g_{32} & g_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ g_{n1} & g_{n2} & g_{n3} & \cdots & \cdots & g_{nn} \end{pmatrix} \begin{pmatrix} g_{11} & g_{21} & g_{31} & \cdots & \cdots & g_{n1} \\ 0 & g_{22} & g_{32} & \cdots & \cdots & g_{n2} \\ 0 & 0 & g_{33} & \cdots & \cdots & g_{n3} \\ \vdots & \vdots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & g_{nn} \end{pmatrix}$$

e,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix}$$

Como existe uma lei de formação para os elementos diagonais e outra para os não diagonais de G, veremos como obter as fórmulas em separado.

a) Elementos diagonais de G

Os elementos diagonais a_{ii} de A são iguais ao produto da linha i de G pela coluna i de G^T. Então este produto é equivalente a se multiplicar a linha i de G por ela mesma. Portanto:

$$\begin{aligned} &a_{11} = g_{11}^2, \\ &a_{22} = g_{21}^2 + g_{22}^2, \\ &\cdots \\ &g_{nn} = g_{n1}^2 g_{n2}^2 + \cdots + g_{nn}. \end{aligned}$$

Assim, os elementos diagonais de G são dados por:

$$\begin{cases} g_{11} = \sqrt{a_{11}}, \\ \\ g_{ii} = \sqrt{a_{ii} - \displaystyle\sum_{k=1}^{i-1} g_{ik}^2} \ , \quad i = 2,\ 3,\ \cdots, n. \end{cases} \tag{2.18}$$

b) Elementos não diagonais de G

1ª coluna : A 1ª coluna de G é obtidos igualando-se os elementos da 1ª coluna de A (abaixo da diagonal principal) com o produto de cada linha de G pela 1ª coluna de G^T.
Portanto:

$$\begin{aligned} &a_{21} = g_{21} \cdot g_{11}, \\ &a_{31} = g_{31} \cdot g_{11}, \\ &\cdots \\ &a_{n1} = g_{n1} \cdot g_{11}. \\ &\textit{ou seja :} \\ &g_{i1} = \frac{a_{i1}}{g_{11}}, \quad i = 2,\ 3,\ \cdots,\ n. \end{aligned}$$

2ª coluna : A 2ª coluna de G é obtidos igualando-se os elementos da 2ª coluna de A (abaixo da diagonal principal) com o produto de cada linha de G pela 2ª coluna de G^T.
Portanto:

$$\begin{aligned} &a_{32} = g_{31} \cdot g_{21} \cdot + g_{32} \cdot g_{22}, \\ &a_{42} = g_{41} \cdot g_{21} \cdot + g_{42} \cdot g_{22}, \\ &\cdots \\ &a_{n2} = g_{n1} \cdot g_{21} \cdot + g_{n2} \cdot g_{22}, \\ &\textit{ou seja :} \\ &g_{i2} = \frac{a_{i2} - g_{i1} \cdot g_{21}}{g_{22}}, \quad i = 3,\ 4,\ \cdots,\ n. \end{aligned}$$

Continuando os cálculos para 3ª, 4ª , etc. ... colunas de G, tem-se a fórmula geral:

$$\begin{cases} g_{i1} = \dfrac{a_{i1}}{g_{11}}, \quad i = 2,\ 3,\ \cdots,\ n. \\ \\ g_{ij} = \dfrac{a_{ij} - \displaystyle\sum_{k=1}^{j-1} g_{ik} \cdot g_{jk}}{g_{jj}}, \quad 2 \leq j < i. \end{cases} \tag{2.19}$$

Quando utilizadas numa ordem adequada as fórmulas (2.18) e (2.19) determinam os elementos da matriz G.
Uma ordem conveniente pode ser:

$$g_{11},\ g_{21},\ g_{31},\ \cdots,\ g_{n1};\ g_{22},\ g_{32},\ \cdots,\ g_{n2};\ \cdots, g_{nn}.$$

Isto significa que calcula-se os elementos da matriz G por coluna.

Nota *Se A satisfaz as condições do método de Cholesky, a aplicação do método requer menos cálculos que as decomposições LU de Crout e Doolittle.*

Nota *A positiva definida garante que na decomposição teremos somente raízes quadradas de números positivos.*

Nota *O método de Cholesky pode também ser aplicado a matrizes simétricas não positivas definidas desde que se use números complexos. Entretanto, só será usado o método de Cholesky com números reais.*

Nota *Pode ser mostrado que se o sistema de equações algébricas $A \cdot \boldsymbol{x} = \boldsymbol{b}$, onde A é uma matriz não singular, é transformado no sistema equivalente $A' \cdot \boldsymbol{x} = \boldsymbol{b'}$, com $A' = A^T \cdot A; \boldsymbol{b'} = A^T \cdot b$, onde A^T é a transposta de A, então o último sistema pode sempre ser resolvido pelo processo de Cholesky (isto é, a matriz A' satisfaz as condições para a aplicação do método).*

Nota Determinantes*:*
Na decomposição LU de Cholesky tem-se que $A = G \cdot G^T$ e portanto:

$$det(A) = [det(G)]^2 = (g_{11} g_{22} \cdots \ g_{nn})^2$$

2.1.3.5 Solução de Sistemas por Cholesky

A resolução de sistemas lineares é semelhante ao método LU. Seja $A = G \cdot G^T$, então resolver $A \cdot \mathbf{x} = \mathbf{b}$ é equivalente a resolver $G \cdot \mathbf{y} = \mathbf{b}$ e depois $G^T \cdot \mathbf{x} = \mathbf{y}$:

$$\begin{cases} G \cdot \mathbf{y} = \mathbf{b} \\ \\ G^T \cdot \mathbf{x} = \mathbf{y} \end{cases}$$

Exemplo 2.7

Seja:

$$A = \begin{bmatrix} 4 & -2 & 2 \\ -2 & 10 & -1 \\ 2 & -1 & 2 \end{bmatrix} \ e \ \mathbf{b} = \begin{bmatrix} 6 \\ 15 \\ 2 \end{bmatrix}$$

a) Decompor A em GG^T.
b) Calcular o determinante de A, usando a decomposição obtida.
c) Resolver o sistema $A \cdot \mathbf{x} = \mathbf{b}$.

Solução *a) Usando as fórmulas (2.18) e (2.19), obtém-se:*

$$g_{11} = \sqrt{a_{11}} \Rightarrow g_{11} = \sqrt{4} \Rightarrow g_{11} = 2,$$

$$g_{21} = \frac{a_{21}}{g_{11}} \Rightarrow g_{21} = \frac{-2}{2} = -1,$$

$$g_{31} = \frac{a_{31}}{g_{11}} \Rightarrow g_{31} = \frac{2}{2} = 1,$$

$$g_{22} = \sqrt{a_{22} - g_{21}^2} \Rightarrow g_{22} = \sqrt{10 - (-1)^2} = 3,$$

$$g_{32} = \frac{a_{32} - g_{31} \cdot g_{21}}{g_{22}} \Rightarrow g_{32} = \frac{-1 - (1 \cdot -1)}{3} = 0,$$

$$g_{33} = \sqrt{a_{33} - g_{31}^2 - g_{32}^2} \Rightarrow g_{33} = \sqrt{2 - 1^2 - 0^2} = 1$$

Temos então:

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

b) $det(A) = (g_{11}g_{22}g_{33})^2 = (231)^2 = 36$.

c) Para obter a solução do sistema $A \cdot \boldsymbol{x} = b$*, deve-se resolver dois sistemas triangulares:* $G \cdot \boldsymbol{y} = \boldsymbol{b}$ *e* $G^T \cdot \boldsymbol{x} = \boldsymbol{y}$.

De $G \cdot \boldsymbol{y} = \boldsymbol{b}$*:*

$$\begin{bmatrix} 2 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 15 \\ 2 \end{bmatrix}$$

$$2 \cdot y_1 = 6 \Rightarrow y_1 = 3$$
$$-y_1 + 3y_2 = 15 \Rightarrow -3 + 3y_2 = 15 \Rightarrow 3y_2 = 15 + 3 \Rightarrow 3y_2 = 18 \Rightarrow y_2 = 6$$
$$y_1 + y_3 = 2 \Rightarrow 3 + y_3 = 2 \Rightarrow y_3 = 2 - 3 \Rightarrow y_3 = -1$$

Então a solução do sistema $G \cdot \boldsymbol{y} = \boldsymbol{b}$ *é* $\boldsymbol{y} = (3,\ 6,\ -1)^T$.

De $G^T \cdot \boldsymbol{x} = \boldsymbol{y}$*:*

$$\begin{bmatrix} 2 & -1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ -1 \end{bmatrix}$$

$$x_3 = -1$$
$$3x_2 = 6 \Rightarrow x_2 = 2$$
$$2x_1 - x_2 + x_3 = 3 \Rightarrow 2x_1 - 2 - 1 = 3 \Rightarrow 2x_1 = 6 \Rightarrow x_1 = 3$$

Então a solução do sistema $G^T \cdot \boldsymbol{x} = \boldsymbol{y}$ *é* $\boldsymbol{x} = (3,\ 2,\ -1)^T$.

Portanto, a solução do sistema $A \cdot \boldsymbol{x} = \boldsymbol{b}$ *, isto é, de:*

$$\begin{bmatrix} 4 & -2 & 2 \\ -2 & 10 & -1 \\ 2 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 15 \\ 2 \end{bmatrix} \quad \acute{e} \quad \begin{bmatrix} 3 \\ 2 \\ -1 \end{bmatrix}$$

2.1.3.6 Fatoração de Choleski : $A = GG^t$ para matrizes simétricas

Exemplo 2.8

Resolver o sistema de equações lineares abaixo pelo método de Choleski:

$$\begin{cases} 3x + 2y + 4z &= 1 \\ x + y - 2z &= 0 \\ 4x + 3y - 2z &= 2 \end{cases}$$

Transfor-

$$A = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & -2 \\ 4 & 3 & -2 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

mando o sistema em matriz simétrica:

$$A \cdot A^t = \begin{pmatrix} 26 & 19 & 2 \\ 19 & 14 & 0 \\ 2 & 0 & 24 \end{pmatrix} \qquad b \cdot b^t = \begin{pmatrix} 11 \\ 8 \\ 0 \end{pmatrix}$$

Fatoração LU da matrix simétrica:

$$\begin{matrix} -\frac{19}{26} = -0.73 \\ -\frac{2}{26} = -0.077 \end{matrix} \begin{bmatrix} 26 & 19 & 2 \\ 19 & 14 & 0 \\ 2 & 0 & 24 \end{bmatrix} = \begin{bmatrix} 26 & 19 & 2 \\ \boxed{0.73} & 0.13 & -1.46 \\ \boxed{0.077} & -1.46 & 24 \end{bmatrix}$$

$$-\frac{-1.46}{0.13} = 11.23 \begin{bmatrix} 26 & 19 & 2 \\ \boxed{0.73} & 0.13 & -1.46 \\ \boxed{0.077} & -1.46 & 24 \end{bmatrix} = \begin{bmatrix} 26 & 19 & 2 \\ \boxed{0.73} & 0.13 & -1.46 \\ \boxed{0.077} & \boxed{\text{-11.23}} & 7.60 \end{bmatrix} = "LU"$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.73 & 1 & 0 \\ 0.077 & -11.23 & 1 \end{bmatrix}; \quad U = \begin{bmatrix} 26 & 19 & 2 \\ 0 & 0.13 & -1.46 \\ 0 & 0 & 7.6 \end{bmatrix}$$

Fatoração LDL^t da decomposição LU:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.73 & 1 & 0 \\ 0.077 & -11.23 & 7.60 \end{bmatrix}; \quad U = \begin{bmatrix} 26 & 19 & 2 \\ 0 & 0.13 & -1.46 \\ 0 & 0 & 7.6 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.73 & 1 & 0 \\ 0.077 & -11.23 & 1 \end{bmatrix}; \quad D = \begin{bmatrix} 26 & 0 & 0 \\ 0 & 0.13 & 0 \\ 0 & 0 & 7.6 \end{bmatrix}; \quad L^t = \begin{bmatrix} 1 & 0.73 & 0.077 \\ 0 & 1 & -11.23 \\ 0 & 0 & 1 \end{bmatrix}$$

$G = L \cdot \sqrt{D}$:

$$G = \begin{bmatrix} 1 & 0 & 0 \\ 0.73 & 1 & 0 \\ 0.077 & -11.23 & 1 \end{bmatrix} \times \begin{bmatrix} 5.099 & 0 & 0 \\ 0 & 0.3606 & 0 \\ 0 & 0 & 2.7568 \end{bmatrix} = \begin{bmatrix} 5.099 & 0 & 0 \\ 3.722 & 0.3606 & 0 \\ 0.3926 & -4.049 & 2.7568 \end{bmatrix}$$

$G^t = \sqrt{D} \cdot L^t$:

$$G^t = \begin{bmatrix} 5.099 & 0 & 0 \\ 0 & 0.3606 & 0 \\ 0 & 0 & 2.7568 \end{bmatrix} \times \begin{bmatrix} 1 & 0.73 & 0.077 \\ 0 & 1 & -11.23 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 5.099 & 3.722 & 0.3926 \\ 0 & 0.3606 & -4.049 \\ 0 & 0 & 2.7568 \end{bmatrix}$$

2.1.3.7 Fatoração de Choleski: $A = GG^t$ Aplicado à Solução de Sistemas Lineares

Partindo da decomposição $A = G \cdot G^t$, teremos a solução através de dois sistemas triangulares:
De $G\mathbf{y} = \mathbf{b}$ obtém-se $\mathbf{y}$
Agora de $G\mathbf{x} = \mathbf{y}$ obtém-se
$\mathbf{x}$ como solução do sistema $A\mathbf{x} = \mathbf{b}$

Execícios 2.1

1) Aplicando-se o processo de Cholesky a matriz A, obteve-se

$$A = \begin{pmatrix} \cdots & 2 & \cdots & \cdots \\ \cdots & 8 & 10 & -8 \\ 3 & 10 & 14 & -5 \\ \cdots & -8 & \cdots & 29 \end{pmatrix} = G \cdot G^T \text{ onde } G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & \cdots & 0 & 0 \\ \cdots & 2 & 1 & 0 \\ 0 & -4 & \cdots & 2 \end{pmatrix}$$

Preencher os espaços pontilhados com valores apropriados.

2) Considere as matrizes:

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & -1 & 3 \end{pmatrix}; \quad B = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 1 \end{pmatrix}$$

Escolha apropriadamente e resolva um dos sistemas : $A \cdot \mathbf{x} = \mathbf{b}$, $B \cdot \mathbf{x} = \mathbf{b}$, e resolva pelo método de Cholesky, onde $b = (2,\ 1,\ 5)^T$.

2.1.4 Armadilhas dos métodos diretos

Todos os sistemas não singulares de equações algébricas lineares têm uma solução. Em teoria, a solução sempre pode ser obtida pela eliminação de Gauss. No entanto, existem duas armadilhas principais na aplicação da eliminação de Gauss (ou suas variações): (a) presença de erros de arredondamentos e (b) sistemas mal-condicionados.

2.2 Métodos Iterativos

Até agora foram apresentados apenas métodos diretos de solução. A característica comum desses métodos é que eles encontram a solução com um número fixo de operações. Além disso, se o computador fosse capaz de realizar cálculos com precisão infinita (sem erros de arredondamento), a solução seria exata. Métodos iterativos ou métodos indiretos começam com uma estimativa inicial da solução **x** e depois repetidamente melhoram a solução até que a mudança nas aproximações $\tilde{\mathbf{x}}$ para **x** tornam-se insignificantes. Os métodos iterativos têm as seguintes vantagens que os tornam atraentes para certos problemas:

1. É possível armazenar somente os elementos diferentes de zero da matriz de coeficientes. Este faz com que seja possível lidar com matrizes muito grandes que são esparsas.
2. Os processos iterativos são auto-corretivos, o que significa que erros de arredondamento (ou mesmo erros aritméticos), em um ciclo iterativo são corrigidos em ciclos seguintes.

Uma vez que o número necessário de iterações pode ser muito grande, os métodos indiretos são, em geral, mais lentos do que os métodos diretos.

Um inconveniente grave dos métodos iterativos é que nem sempre convergem para a solução. Pode ser mostrado que a convergência é garantida se a matriz de coeficientes é diagonalmente dominante. A estimativa inicial de x não desempenha nenhum papel na determinação se a convergência ocorre - se o processo converge para um vetor inicial, ele convergirá qualquer vetor inicial. O vetor inicial afeta apenas o número de iterações que são necessárias para a convergência.

Seja M uma matriz chamada de matriz de iteração e **c** um vetor constante. Um método iterativo escrito na forma

$$\mathbf{x}^{[k]} = M \cdot \mathbf{x}^{[k-1]} + \mathbf{c}, \quad k = 1, 2, \cdots, n \tag{2.20}$$

é chamado de estacionário quando a matriz M for fixa durante o processo de iteração. Na próxima seção serão apresentados dois métodos estacionários: Jacobi e Gauss-Seidel.

2.2.1 Teste de Parada

Como em todos os processos iterativos, necessita-se de um critério para a parada do processo.

a) Máximo desvio absoluto:

$$\delta^{[k]} = \max_{i=1,n} \left| x_i^{[k]} - x_i^{[k-1]} \right| \tag{2.21}$$

b) Máximo desvio relativo:

$$\delta_R^{[k]} = \frac{\delta^{[k]}}{\max\limits_{i=1,n} \left| x_i^{[k]} \right|} \tag{2.22}$$

c) Número máximo de iterações:

$$k \geq k_{max} \tag{2.23}$$

Desta forma, dada uma precisão ε, o vetor $\tilde{x}^{[k]}$ será escolhido como solução aproximada da solução exata, se $\delta^{[k]} < \varepsilon$, ou dependendo da escolha, $\delta_R^{[k]} < \varepsilon$. No caso em $k \geq k_{max}$, $\tilde{x}^{[k_{max}]}$ será escolhido como solução aproximada da solução exata.

2.3 Métodos Iterativos Estacionários

Consideremos um sistema genérico $A \cdot \mathbf{x} = \mathbf{b}$ escrito na forma $A = M - N$. Supondo que M tem inversa, obtém-se

$$(M - N) \cdot \mathbf{x} = \mathbf{b} \quad M \cdot \mathbf{x} = \mathbf{b} + N \cdot \mathbf{x} \quad \mathbf{x} = M^{-1}(\mathbf{b} + N \cdot \mathbf{x})$$

Agora podemos definir um método iterativo que consiste em:

Escolher um vetor inicial $\mathbf{x}^{[0]}$
Iteração $\mathbf{x}^{[k]} = M^{-1}(b + N\mathbf{x}^{[k-1]})$ $(k = 1,\ 2,\ \cdots n)$

É importante que a matriz M seja muito mais simples do que A, porque senão estaríamos complicando o problema.

Se $\left\| M^{-1} \cdot N \right\| < 1$, a sequência definida pela iteração $\mathbf{x}^{[k]} = M^{-1}(b + N\mathbf{x}^{[k-1]})$, $(k = 1,\ 2,\ \cdots n)$ converge para o ponto fixo do sistema de equações, qualquer que seja $\mathbf{x}^{[0]} \in \mathbb{R}$. A taxa de convergência deste método iterativo é linear e a constante de convergência é menor ou igual a $\left\| M^{-1} \cdot N \right\|$. Diferentes escolhas de M e N definem diferentes métodos iterativos. Considere-se a seguinte decomposição da matriz A na soma de três matrizes $A = L + D + U$,

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & 0 \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{2.24}$$

isto é, $A = L + D + U$, onde L é a matriz triangular inferior, U a matriz triangular superior, ambas com zeros na diagonal principal e D a matriz diagonal. Nota-se que a matriz diagonal D não deverá ter zeros na diagonal principal. Caso isso aconteça, deve-se efetuar uma troca de linhas ou colunas na matriz A, para obtermos uma matriz D adequada.

Dentre os métodos iterativos estacionários, iremos abordar os seguintes:

1. Método de Jacobi;
2. Método de Gauss-Seidel.

2.3.1 Método de Jacobi

Vamos supor que A seja ordenada de modo que todos os seus elementos da diagonal sejam não-nulos $a_{ii} \neq 0,\ \forall i \leq n$. No caso do método de Jacobi, consideramos $A = D + L + U = M - N$

$$M = D$$
$$N = -(L + U)$$

Portanto, o método consiste em

Escolher um vetor inicial $\mathbf{x}^{[0]}$
Iteração $\mathbf{x}^{[k]} = D^{-1} \cdot \left(\mathbf{b} + N \cdot \mathbf{x}^{[k-1]}\right)$ $(k = 1,\ 2,\ \cdots n)$

ou ainda

$$x_i^{[k]} = \frac{1}{a_{ii}} \cdot \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} \cdot x_j^{[k-1]} \right), \quad (i, k = 1,\ 2,\ \cdots,\ n) \tag{2.25}$$

Na prática vamos então explicitar o valor de cada x_i na $i-$ésima equação $(i = 1,\ 2,\ \cdots,\ n)$. Como se assume que a_{ii} é não nulo, pode-se escrever:

$$\begin{aligned}
x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n) \\
x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n) \\
x_3 &= \frac{1}{a_{33}}(b_3 - a_{31}x_1 - a_{32}x_2 - \cdots - a_{3n}x_n) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
x_n &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n\,n-1}x_{n-1})
\end{aligned} \tag{2.26}$$

Se considerarmos o lado esquerdo do sistema como os elementos de um novo passo de iteração $[k]$ e os elementos do lado direito como elementos do passo anterior $[k-1]$, teremos:

$$\begin{aligned}
x_1^{[k]} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^{[k-1]} - a_{13}x_3^{[k-1]} - \cdots - a_{1n}x_n^{[k-1]}) \\
x_2^{[k]} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{[k-1]} - a_{23}x_3^{[k-1]} - \cdots - a_{2n}x_n^{[k-1]}) \\
x_3^{[k]} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^{[k-1]} - a_{32}x_2^{[k-1]} - \cdots - a_{3n}x_n^{[k-1]}) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
x_n^{[k]} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{[k-1]} - a_{n2}x_2^{[k-1]} - \cdots - a_{n,\,n-1}x_{n-1}^{[k-1]})
\end{aligned} \tag{2.27}$$

Exemplo 2.9

Resolver o sistema abaixo pelo método de Jacobi.

$$\begin{cases} 30x - 10y - 2z = 78 \\ x + 70y - 3z = -120 \\ 3x - 2y + 100z = 900 \end{cases}$$

O sistema acima produz as seguintes equações do procedimento iterativo

$$x^{[k]} = \frac{78 + 10y^{[k-1]} + 2z^{[k-1]}}{30}$$

$$y^{[k]} = \frac{-120 - x^{[k-1]} + 3z^{[k-1]}}{70}$$

$$z^{[k]} = \frac{900 - 3x^{[k-1]} + 2y^{[k-1]}}{100}$$

Assumindo

$$\begin{aligned} \mathbf{x}^{[0]} &= (x^{[0]}, y^{[0]}, z^{[0]}) \\ &= (0\,, 0\,, 0\,) \end{aligned}$$

Realiza-se a 1ª iteração

$$x^{[1]} = \frac{78 + 10y^{[0]} + 2z^{[0]}}{30} = \frac{78 + 10 \times 0 + 2 \times 0}{30} = 2.6$$

$$y^{[1]} = \frac{-120 - x^{[0]} + 3z^{[0]}}{70} = \frac{-120 - 0 + 3 \times 0}{70} = -1.71$$

$$z^{[1]} = \frac{900 - 3x^{[0]} + 2y^{[0]}}{100} = \frac{900 - 3 \times 0 + 2 \times 0}{100} = 9$$

$$\begin{aligned} Erro &= \max(\ |x^{[1]} - x^{[0]}|,\ |y^{[1]} - y^{[0]}|,\ |z^{[1]} - z^{[0]}|) \\ Erro &= \max(\ |2.6 - 0|,\ |-1.71 - 0|,\ |9 - 0|) = 9 \end{aligned}$$

Realiza-se a 2ª iteração

$$x^{[2]} = \frac{78 + 10y^{[1]} + 2z^{[1]}}{30} = \frac{78 + 10 \times -1.71 + 2 \times 9}{30} = 2.63$$

$$y^{[2]} = \frac{-120 - x^{[1]} + 3z^{[1]}}{70} = \frac{-120 - 2.6 + 3 \times 9}{70} = -1.37$$

$$z^{[2]} = \frac{900 - 3x^{[1]} + 2y^{[1]}}{100} = \frac{900 - 3 \times 2.6 + 2 \times -1.71}{100} = 8.89$$

$$\begin{aligned} Erro &= \max(\ |x^{[2]} - x^{[1]}|,\ |y^{[2]} - y^{[1]}|,\ |z^{[2]} - z^{[1]}|) \\ Erro &= \max(\ |2.63 - 2.6|,\ |-1.37 - (-1.71)|,\ |8.89 - 9|) = 0.34 \end{aligned}$$

Realiza-se a 3ª iteração

$$x^{[3]} = \frac{78 + 10y^{[2]} + 2z^{[2]}}{30} = \frac{78 + 10 \times -1.37 + 2 \times 8.89}{30} = 2.74$$

$$y^{[3]} = \frac{-120 - x^{[2]} + 3z^{[2]}}{70} = \frac{-120 - 2.63 + 3 \times 8.89}{70} = -1.37$$

$$z^{[3]} = \frac{900 - 3x^{[2]} + 2y^{[2]}}{100} = \frac{900 - 3 \times 2.63 + 2 \times -1.37}{100} = 8.89$$

$$Erro = \max(\ |x^{[3]} - x^{[2]}|,\ |y^{[3]} - y^{[2]}|,\ |z^{[3]} - z^{[2]}|)$$
$$Erro = \max(\ |2.74 - 2.63|,\ |-1.37 - (-1.37)|,\ |8.89 - 8.89|) = 0.11$$

Realiza-se a 4ª iteração

$$x^{[4]} = \frac{78 + 10y^{[3]} + 2z^{[3]}}{30} = \frac{78 + 10 \times -1.37 + 2 \times 8.89}{30} = 2.74$$

$$y^{[4]} = \frac{-120 - x^{[3]} + 3z^{[3]}}{70} = \frac{-120 - 2.74 + 3 \times 8.89}{70} = -1.37$$

$$z^{[4]} = \frac{900 - 3x^{[3]} + 2y^{[3]}}{100} = \frac{900 - 3 \times 2.74 + 2 \times -1.37}{100} = 8.89$$

$$Erro = \max(\ |x^{[4]} - x^{[3]}|,\ |y^{[4]} - y^{[3]}|,\ |z^{[4]} - z^{[3]}|)$$
$$Erro = \max(\ |2.74 - 2.74|,\ |-1.37 - (-1.37)|,\ |8.89 - 8.89|) < 0.01$$

Pode-se exibir todas iterações numa tabela resumo:

Tabela 2.2: Tabela resumo do Método de Jacobi.

k	$x^{[k]}$	$y^{[k]}$	$z^{[k]}$	$Erro^{[k]}$
0	0	0	0	-
1	2.6	-1.71	9	9
2	2.63	-1.37	8.89	0.34
3	2.74	-1.37	8.89	0.11
4	2.74	-1.37	8.89	< 0.01

2.3.2 Método de Gauss-Seidel

No caso do método de Gauss-Seidel, poderemos considerar $A = D + L + U = M - N$

$$M = D + L$$
$$N = -U$$

Portanto o método consiste em

Escolher um vetor inicial $\mathbf{x}^{[0]}$
Iteração $\mathbf{x}^{[k]} = (D + L)^{-1} \cdot \left(\mathbf{b} \ - U \cdot \mathbf{x}^{[k-1]}\right)$ $(k = 1,\ 2,\ \cdots n)$

O princípio do método de Gauss-Seidel é usar nova informação tão logo ela esteja disponível. Então:

$$(D+L)\cdot \mathbf{x}^{[k]} = [\mathbf{b} \ -U\cdot \mathbf{x}^{[k-1]}], \quad (k=1,\ 2,\ \cdots,\ n)$$

$$\mathbf{x}^{[k]} = D^{-1}\cdot [\mathbf{b} - L\cdot \mathbf{x}^{[k]} - U\cdot \mathbf{x}^{[k-1]}], \quad (k=1,\ 2,\ \cdots,\ n)$$

ou seja

Escolher um vetor inicial $\mathbf{x}^{(0)}$
Iteração $\mathbf{x}^{[k]} = D^{-1}\cdot [\mathbf{b} - L\cdot \mathbf{x}^{[k]} - U\cdot \mathbf{x}^{[k-1]}]$ $(k=1,\ 2,\ \cdots n)$

ou ainda

$$\begin{array}{l} x_i^{[k]} = \frac{1}{a_{ii}}\cdot \left(b_i - \sum_{j=1}^{i-1} a_{ij}\cdot x_j^{[k-1]} - \sum_{j=i+1}^{n} a_{ij}\cdot x_j^{(k)}\right), \\ (i=1,\ 2,\ \cdots,\ n) \end{array} \tag{2.28}$$

Se considerarmos o lado esquerdo do sistema como os elementos de um novo passo de iteração $[k]$ e os elementos do lado direito como elementos novos tão logo eles estejam disponíveis, tem-se:

$$\begin{array}{rcl} x_1^{[k]} & = & \frac{1}{a_{11}}(b_1 - a_{12}x_2^{[k-1]} - a_{13}x_3^{[k-1]} - \cdots - a_{1n}x_n^{[k-1]}) \\ x_2^{[k]} & = & \frac{1}{a_{22}}(b_2 - a_{21}x_1^{[k]} - a_{23}x_3^{[k-1]} - \cdots - a_{2n}x_n^{[k-1]}) \\ x_3^{[k]} & = & \frac{1}{a_{33}}(b_3 - a_{31}x_1^{[k]} - a_{32}x_2^{[k]} - \cdots - a_{3n}x_n^{[k-1]}) \\ \vdots & = & \vdots \qquad\qquad \vdots \\ x_n^{[k]} & = & \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{[k]} - a_{n2}x_2^{[k]} - \cdots - a_{n,\,n-1}x_{n-1}^{[k]}) \end{array} \tag{2.29}$$

Exemplo 2.10

Resolver o sistema abaixo pelo método de Gauss-Seidel.

$$\begin{cases} 30x - 10y - 2z = 78 \\ x + 70y - 3z = -120 \\ 3x - 2y + 100z = 900 \end{cases}$$

O sistema acima produz as seguintes equações do procedimento iterativo

$$x^{[k]} = \frac{78 + 10y^{[k-1]} + 2z^{[k-1]}}{30}$$

$$y^{[k]} = \frac{-120 - x^{[k]} + 3z^{[k-1]}}{70}$$

$$z^{[k]} = \frac{900 - 3x^{[k]} + 2y^{[k]}}{100}$$

$$k = 1, 2\cdots$$

Assumindo

$$\begin{aligned}\mathbf{x}^{[0]} &= (x^{[0]}, y^{[0]}, z^{[0]}) \\ &= (0\,, 0\,, 0\,)\end{aligned}$$

Realiza-se a 1ª iteração

$$x^{[1]} = \frac{78 + 10y^{[0]} + 2z^{[0]}}{30} = \frac{78 + 10 \times 0 + 2 \times 0}{30} = 2.6$$

$$y^{[1]} = \frac{-120 - x^{[1]} + 3z^{[0]}}{70} = \frac{-120 - 2.6 - 3 \times 0}{70} = -1.75$$

$$z^{[1]} = \frac{900 - 3x^{[1]} + 2y^{[1]}}{100} = \frac{900 - 3 \times 2.6 + 2 \times -1.75}{100} = 8.89$$

$$\begin{aligned}Erro &= \max(\ |x^{[1]} - x^{[0]}|,\ |y^{[1]} - y^{[0]}|,\ |z^{[1]} - z^{[0]}|) \\ Erro &= \max(\ |2.6 - 0|,\ |-1.75 - 0|,\ |8.89 - 0|) = 8.89\end{aligned}$$

Realiza-se a 2ª iteração

$$x^{[2]} = \frac{78 + 10y^{[1]} + 2z^{[1]}}{30} = \frac{78 + 10 \times -1.75 + 2 \times 8.89}{30} = 2.61$$

$$y^{[2]} = \frac{-120 - x^{[2]} + 3z^{[1]}}{70} = \frac{-120 - 2.61 - 3 \times 8.89}{70} = -1.37$$

$$z^{[2]} = \frac{900 - 3x^{[2]} + 2y^{[2]}}{100} = \frac{900 - 3 \times 2.61 + 2 \times -1.37}{100} = 8.89$$

$$\begin{aligned}Erro &= \max(\ |x^{[2]} - x^{[1]}|,\ |y^{[2]} - y^{[1]}|,\ |z^{[2]} - z^{[1]}|) \\ Erro &= \max(\ |2.61 - 2.6|,\ |-1.37 - (-1.75)|,\ |8.89 - 8.89|) = 0.38\end{aligned}$$

Realiza-se a 3ª iteração

$$x^{[3]} = \frac{78 + 10y^{[2]} + 2z^{[2]}}{30} = \frac{78 + 10 \times -1.37 + 2 \times 8.89}{30} = 2.74$$

$$y^{[3]} = \frac{-120 - x^{[3]} + 3z^{[2]}}{70} = \frac{-120 - 2.74 - 3 \times 8.89}{70} = -1.37$$

$$z^{[3]} = \frac{900 - 3x^{[3]} + 2y^{[3]}}{100} = \frac{900 - 3 \times 2.74 + 2 \times -1.37}{100} = 8.89$$

$$\begin{aligned}Erro &= \max(\ |x^{[3]} - x^{[2]}|,\ |y^{[3]} - y^{[2]}|,\ |z^{[3]} - z^{[2]}|) \\ Erro &= \max(\ |2.74 - 2.61|,\ |-1.37 - (-1.37)|,\ |8.89 - 8.89|) = 0.13\end{aligned}$$

Realiza-se a 4ª iteração

$$\begin{aligned}x^{[4]} &= \frac{78 + 10y^{[3]} + 2z^{[3]}}{30} = \frac{78 + 10 \times -1.37 + 2 \times 8.89}{30} = 2.74 \\ y^{[4]} &= \frac{-120 - x^{[4]} + 3z^{[3]}}{70} = \frac{-120 - 2.74 - 3 \times 8.89}{70} = -1.37 \\ z^{[4]} &= \frac{900 - 3x^{[4]} + 2y^{[4]}}{100} = \frac{900 - 3 \times 2.74 + 2 \times -1.37}{100} = 8.89 \\ Erro &= \max(\ |x^{[4]} - x^{[3]}|,\ |y^{[4]} - y^{[3]}|,\ |z^{[4]} - z^{[3]}|) \\ Erro &= \max(\ |2.74 - 2.74|,\ |-1.37 - (-1.37)|,\ |8.89 - 8.89|) < 0.01\end{aligned}$$

Pode-se exibir todas iterações numa tabela resumo:

Tabela 2.3: Tabela resumo do Método de Gauss-Seidel.

k	$x^{[k]}$	$y^{[k]}$	$z^{[k]}$	$Erro^{[k]}$
0	0	0	0	-
1	2.6	-1.75	8.89	8.89
2	2.61	-1.37	8.89	0.38
3	2.74	-1.37	8.89	0.13
4	2.74	-1.37	8.89	< 0.01

2.4 Convergência dos Métodos Iterativos

Observa-se que o método iterativo $\mathbf{x}^{[k]} = M^{-1} \cdot \left(\mathbf{b} + N \cdot \mathbf{x}^{[k]}\right)$ pode ser escrito como:

$$\mathbf{x}^{[k]} = C \cdot \mathbf{x}^{[k-1]} + \mathbf{d}, \quad k = 1, 2, \cdots$$

onde $C = M^{-1} \cdot N$ e $\mathbf{d} = M^{-1} \cdot \mathbf{b}$.

Teorema 2.2

O método iterativo $\boldsymbol{x}^{[k]} = C \cdot \boldsymbol{x}^{[k-1]} + \boldsymbol{d}$ converge com qualquer vator inicial $\boldsymbol{x}^{[0]}$ se, e somente se, $\rho(C) < 1$, sendo $\rho(C)$ o raio espectral (maior autovalor em módulo, isto é, $\rho(C) = max|\lambda|$, onde λ é um autovalor de C da matriz de iteração C).

Lembrando que:

1. Método de Jacobi: $C = -D^{-1} \cdot (L + U)$;
2. Método de Gauss-Seidel: $C = -(L + D)^{-1} \cdot U$;

A determinação do raio espectral da matriz de iteração (C) requer, em geral, maior esforço computacional que a pró pia solução do sistema Ax = b. Por isto usa-se normalmente condições suficientes de convergência.

Nota Se existir uma norma induzida $||.|| : ||C|| < 1$ então isso irá se verificar, porque $||\mathbf{e}^{(k)}|| = ||C^k \cdot e^{(0)}|| < ||C||^k ||e^{(0)}|| \longrightarrow 0$, quando k tende para infinito, qualquer que seja o vetor inicial $e^{(0)}$.

Nota Pode-se falar também de ordem de convergência, neste caso vetorial, e as majorações revelam que estes métodos iterativos têm uma convergência linear.

Exemplo 2.11 Verificar se o sistema abaixo pode ser resolvido pelos métodos de Jacobi ou Gauss-Seidel:

$$\begin{pmatrix} 2 & 3 & 1 \\ 1 & 1 & 1 \\ 4 & -4 & 10 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ -3 \end{pmatrix}$$

O roteiro Scilab abaixo encontra os raios espectrais das matrizes de iteração dos métodos de Jacobi e Gauss-Seidel.

```
1.  D=diag([2,1,10]);
2.  L=[ 0  0  0
3.      1  0  0
4.      4 -4  0];
5.  U=[ 0  3  1
6.      0  0  1
7.      0  0  0];
8.  A=D+L+U
9.  J=-inv(D)*(L+U)
10. x=spec(J)
11. rhoJ=max(abs(x))
12. S=inv(-(L+D))*U
13. x=spec(S)
14. rhoS=max(abs(x))
```

Ao executar o roteiro pode-se realizar obter os raios espectrais. Aqui estão os resultados:

```
-->D=diag([2,1,10]);

-->L=[ 0  0  0
-->    1  0  0
-->    4 -4  0];

-->U=[ 0  3  1
-->    0  0  1
-->    0  0  0];

-->A=D+L+U
 A  =
    2.    3.    1.
    1.    1.    1.
    4.  - 4.    10.

-->J=-inv(D)*(L+U)
 J  =

    0.   - 1.5  - 0.5
  - 1.     0.   - 1.
  - 0.4    0.4    0.

-->x=spec(J)
 x  =

  - 1.2707389
    0.9335588
    0.3371801

-->rhoJ=max(abs(x))
 rhoJ  =

    1.2707389

-->S=inv(-(L+D))*U
 S  =

    0.  - 1.5  - 0.5
    0.    1.5  - 0.5
    0.    1.2    0.

-->x=spec(S)
 x  =

    0
    0.75 + 0.1936492i
    0.75 - 0.1936492i

-->rhoS=max(abs(x))
 rhoS  =

    0.7745967

-->
```

Como $\rho(S) < 1$ e $\rho(J) > 1$, então o método de Gauss-Seidel converge e o de Jacobi não. As iterações abaixo confirmam:

A solução pelo método de Jacobi (Diverge):

k	$x^{[k]}$	$y^{[k]}$	$z^{[k]}$	$Erro^{[k]}$
0	0	0	0	-
1	2.5	1	-0.3	2.5
2	1.15	-1.2	-0.9	2.2
3	4.75	0.75	-1.24	3.6
4	1.995	-2.51	-1.9	3.26
5	7.215	0.905	-2.102	5.22
6	2.193	-4.113	-2.824	5.022
7	10.082	1.631	-2.823	7.888
8	1.466	-6.259	-3.68	8.616
9	13.729	3.215	-3.39	12.263
10	-0.627	-9.339	-4.505	14.356

A solução pelo método de Gauss-Seidel (Erro ¡ 0.001):

k	$x^{[k]}$	$y^{[k]}$	$z^{[k]}$	$Erro^{[k]}$
0.	0	0	0	-
1.	2.5	- 1.5	- 1.9	2.5
2.	5.7	- 2.8	- 3.7	3.2
3.	8.55	- 3.85	- 5.26	2.85
4.	10.905	- 4.645	- 6.52	2.355
⋮	⋮	⋮	⋮	⋮
30.	14.996392	- 5.4995935	- 8.4983942	0.0030576
31.	14.998587	- 5.5001932	- 8.4995122	0.0021954
32.	15.000046	- 5.5005337	- 8.5002318	0.0014585
33.	15.000916	- 5.5006846	- 8.5006404	0.0008705

2.4.1 Critérios Suficientes de Convergência

Além do teorema, que nos dá condições necessárias e suficientes de convergência, existem critérios mais simples que asseguram a convergência para qualquer vetor inicial. No entanto, essas condições, que iremos deduzir, são apenas condições suficientes.

No método de Jacobi $C = D^{-1} \cdot (L + U)$:

$$c_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & se \quad i \neq j \\ \\ 0 & se \quad i = j \end{cases} \tag{2.30}$$

e da definição da norma

$$||C||_{\infty} = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |c_{ij}| \tag{2.31}$$

ao exigirmos que a norma do máximo seja inferior a 1, tem-se que

$$\max_{1 \leq i \leq n} \sum_{j=1}^{n} \left| -\frac{a_{ij}}{a_{ii}} \right| < 1 \tag{2.32}$$

Logo, uma condição suficiente que nos garante o método de Jacobi converge, é

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\,, \quad (i = 1, \cdots n) \tag{2.33}$$

neste caso, diz-se que a matriz A tem diagonal estritamente dominante por linhas.
De maneira análoga (usando uma norma semelhante para as colunas), podemos concluir que se

$$|a_{jj}| > \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|\,, \quad (j = 1, \cdots n) \tag{2.34}$$

isto é, se a matriz A tem diagonal estritamente dominante por colunas, então o método de Jacobi converge.

Teorema 2.3

Se a matriz A tiver a diagonal estritamente dominante por linhas ou por colunas, os métodos de Jacobi e de Gauss-Seidel convergem, para qualquer vetor inicial $x^{(0)}$ escolhido. ♡

Nota Como $C = M^{-1} \cdot N = M^{-1}(M - A) = I - M^{-1} \cdot A$ quanto mais próxima de A for a matriz M, mais próximo da matriz zero será valor de C, e consequentemente, mais rápida será a convergência do método iterativo.

M está "mais próxima"de A no caso do Método de Gauss-Seidel ($M = D + L$) do que no caso do Método de Jacobi ($M = D$).

Portanto, habitualmente o método de Gauss-Seidel converge mais rapidamente que o de Jacobi. No entanto tem casos em que isso não acontece, e além disso, um método pode convergir e o outro não!

Nota *Número de Operações* A menos que as matrizes possuam zonas apreciáveis com elementos nulos, ambos os métodos iterativos exigem um cálculo total de $2n^2$ operações, por cada iteração, o que implica que, se forem necessárias mais que $n/3$ iterações, tem-se mais operações do que num método direto.

Capítulo Execícios

1. Resolver o sistema de equações lineares abaixo pelo método de eliminação de Gauss:

$$\begin{cases} 5x - 0.2y - 0.4z &= 4.33 \\ 0.5x + 8y - 0.2z &= -7.3 \\ 0.4x - 0.3y + 12z &= 70 \end{cases}$$

Atenção:

(a). Trabalhar com a matriz de coeficientes ampliada com o vetor constante;
(b). usar 02 casas decimais e no mínimo com 02 dígitos significativos;
(c). exibir cálculos detalhados das etapas de eliminação e substituição retroativa.

2. Resolver o sistema de equações lineares abaixo pelo método de eliminação de Jordan:

$$\begin{cases} 6x - 0.2y - 0.4z &= 9.83 \\ 0.1x + 3y - 0.2z &= -17.3 \\ 0.4x - 0.3y + 7z &= 7 \end{cases}$$

Atenção:

(a). Trabalhar com a matriz de coeficientes ampliada com o vetor constante;
(b). usar 02 casas decimais e no mínimo com 02 dígitos significativos;
(c). exibir cáculos detalhados das etapas de eliminação, substituição retroativa e valores de x, y e z encontrados.

3. Resolver o sistema de equações lineares abaixo pelo método de Jacobi. Utilizar $\mathbf{x} = [0, 0, 0]^t$ como aproximação inicial. Como critério de parada use $Erro^{[k]} = max|x_i^{[k]} - x_i^{[k-1]}| \leq 0.05,\ i = 1..n$. Faça tabela resumo exibindo o erro de cada iteração.

$$\begin{cases} 7x - 2y - z &= 10 \\ 2x - 6y + 2z &= -20 \\ x - 2y + 5z &= -30 \end{cases}$$

Atenção:

(a). usar 02 casas decimais e no mínimo com 02 dígitos significativos;
(b). exibir cálculos detalhados dos valores de $x^{[k]}$, $y^{[k]}$, $z^{[k]}$ e $Erro^{[k]}$ encontrados nas iterações;
(c). faça tabela resumo como exibindo iteração, valores das componentes nas aproximações do vetor solução e estimativa do erro: $Erro^{[k]} = max|x_i^{[k]} - x_i^{[k-1]}|,\ i = 1..n$.

4. Resolver o sistema de equações lineares abaixo pelo método de Gauss-Seidel. Utilizar $\mathbf{x} = [0, 0, 0]^t$ como aproximação inicial. Como critério de parada use $Erro^{[k]} = max|x_i^{[k]} - x_i^{[k-1]}| \leq 0.05,\ i = 1..n$. Faça tabela resumo exibindo o erro de cada iteração.

$$\begin{cases} 5x - 2y - z &= 10 \\ 2x - 6y + 2z &= 20 \\ x - 2y + 16z &= 30 \end{cases}$$

Atenção:

(a). usar 02 casas decimais e no mínimo com 02 dígitos significativos;
(b). exibir cálculos detalhados dos valores de $x^{[k]}$, $y^{[k]}$, $z^{[k]}$ e $Erro^{[k]}$ encontrados nas iterações;
(c). faça tabela resumo como exibindo iteração, valores das componentes nas aproximações do vetor solução e estimativa do erro: $Erro^{[k]} = max|x_i^{[k]} - x_i^{[k-1]}|,\ i = 1..n.$

5. Aplique os critérios de linha e coluna e verifique se os métodos numéricos de Jacobi e Gauss-Seidel tem convergência garantida.

$$a) \begin{cases} 4x + 2y - 9z &= 7 \\ 5x - 6y - 8z &= 3 \\ x - 2y + 15z &= 5 \end{cases}$$

$$b) \begin{cases} 20x + 7y + 9z &= 16 \\ 7x + 30y + 8z &= 38 \\ 19x - 8y + 40z &= 35 \end{cases}$$

$$c) \begin{cases} 20x + 7y + 9z &= 16 \\ 7x + 30y + 8z &= 38 \\ 9x - 18y + 20z &= 35 \end{cases}$$

$$d) \begin{cases} x - 3y + z &= 1 \\ 4x - 18y + 6z &= 3 \\ -x + 3y - z &= 5 \end{cases}$$

$$e) \begin{cases} 4x - y &= 1 \\ -x + 4y - z &= 1 \\ -y + 4z - w &= 1 \\ -z + 4w &= 1 \end{cases}$$

Experimento Numérico:

(a). Resolver os sistemas acima usando os métodos de Jacobi e Guass-Seidel;
(b). use o link para solução on-line `https://atozmath.com/CONM/GaussEli.aspx?q=GJ2` ou;
(c). use o link `http://www.matematica.pucminas.br/lcn/vcn1.htm` e baixe o aplicativo **vcn.exe** para Windows.

Capítulo - Raízes de Equações

3.1 Introdução

Dada uma função $f : \mathbb{R} \to \mathbb{R}$, buscamos um ponto $\xi \in \mathbb{R}$ tal que $f(\xi) = 0$. Este ξ é chamado de uma raiz da equação $f(x) = 0$, ou simplesmente um zero de $f(x)$. No início, só é exigido que $f(x)$ seja contínua no intervalo $[a, b]$ da reta real, $f(x) \in C[a, b]$, e que este intervalo contém a raiz de interesse. A função f poderia ter muitas raízes diferentes; só vamos procurar um. A função $f(x)$ pode ser dada explicitamente, como, por exemplo, um polinômio ou uma função transcendental. Em casos raros, pode ser possível obter as raízes exactas da equação, tal como no caso polinomial favorável. Em geral, no entanto, pode-se esperar obter apenas soluções aproximadas, contando com algum método numérico para produzir a aproximação, de modo que procura-se algoritmos que irá produzir uma solução que é exata com alta precisão, mantendo um mínimo de avaliações de $f(x)$.

Propriedades das Funções Continuas

Definição 3.1

Uma função real é contínua num ponto a se ela é definida em $x = a$ e

$$\lim_{x \to a} f(x) = f(a),$$

isto é, se para todo $\epsilon > 0$ existe um $\delta(\epsilon) > 0$ tal que $|f(x) - f(a)| < \epsilon$ sempre que $|x - a| < \delta(\epsilon)$.

Portanto, se uma função muda gradativamente com as mudanças da variável independente, tal que em cada valor a, da variável independente, a diferença entre $f(x)$ e $f(a)$ aproxima-se de zero quando x aproxima-se de a.

Então, uma função é contínua no ponto a se ambos limites unilaterais (limites direito e esquerdo) são iguais,

$$\lim_{x \to a^-} f(x) = f(a) \;\; e \;\; \lim_{x \to a^+} f(x) = f(a)$$

isto é, se ela é contínua a direita e a esquerda naquele ponto.

Um ponto no qual a o valor da função não é igual ao seu limite, quando x se aproxima deste ponto, é chamado de ponto de descontinuidade. Um função que possui pontos de descontinuidade é descontínua. A função é dita ser contínua se ela é continua em todos os pontos.

Teorema 3.1 (Teorema do valor médio, também conhecido como Teorema de Lagrange)

Suponha uma função $f(x)$ e sua derivada $f'(x)$ são ambas contínuas num intervalo fechado $[a, b]$. Então existe, pelo menos, um ponto $\xi \in [a, b]$ tal que

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

Geometricamente, isto significa que a tangente ao gráfico de $f(x)$ no ponto de abcissa ξ é paralela à secante que passa pelos pontos de abcissas a e b, como mostra a figura (3.1)

O teorema do valor médio também tem uma interpretação em termos físicos: se um objeto está em movimento e se a sua velocidade média é v, então, durante esse percurso no intervalo $[a, b]$, há um instante (ponto ξ) em que a velocidade instantânea também é v.

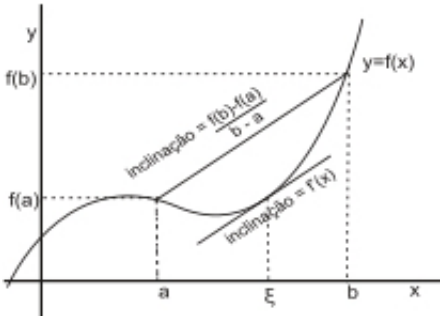


Figura 3.1: Interpretação geométrica do teorema do valor médio

Teorema 3.2 (Teorema do valor intermediário - Teorema de Bolzano)

Uma função real contínua num intervalo fechado $[a, b]$ *então ela tem todos valores entre* $f(a)$ *e* $f(b)$ *para no mínimo um argumento*
Isto é, para todo y *entre* $f(a)$ *e* $f(b)$ *existe pelo menos argumento* c *entre* a *e* b *onde o valor da função* $f(c) = y$*, como mostra a figura (3.2)*

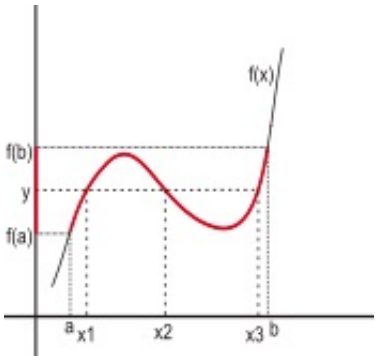


Figura 3.2: Gráfico ilustrativo do teorema do valor intermediário. A função tem três valores de onde $y = f(x), f(a) < y < f(b)$, *i.e.*, $y = f(x_1) = f(x_2) = f(x_3)$.

Teorema 3.3 (Teorema do Valor Extremo)

Seja $f(x)$ *uma função real contínua num intervalo fechado* $[a, b]$ *com* $f(a) \cdot f(b) < 0$*, então existe* x_{min} *e* $x_{max} \in [a, b]$ *tal que para todo* $x \in [a, b]$ *os valores* $f(x_{min}) \leq f(x) \leq f(x_{max})$*. Ver figura (3.3)*

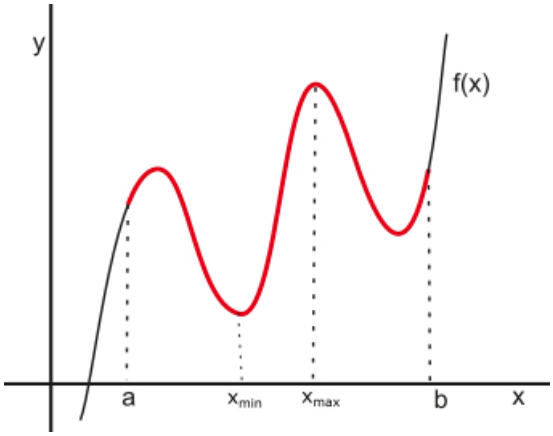


Figura 3.3: Gráfico ilustrativo do teorema do valor extremo.

Teorema 3.4 (Existência de Raízes - Resultado do teorema de Bolzano)

Se $f(x)$ é uma função real contínua num intervalo fechado $[a,b]$ com $f(a)\cdot f(b)<0$, então $\exists\,\xi\in[a,b]$ tal que $f(\xi)=0$.

Então a função $f(x)$ tem pelo menos uma raiz real. Por exemplo, dentro do intervalo $[a,b]$, a função mostrada no gráfico (3.4) tem três raízes, x_1, x_2 e x_3, onde $f(x_1)=0$, $f(x_2)=0$, $f(x_3)=0$, uma vez que $f(a)\cdot f(b)<0$

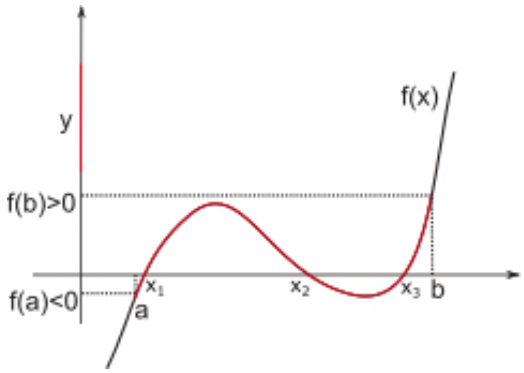


Figura 3.4: Gráfico ilustrativo da existência de raízes. No intervalo $[a,b]$ onde $f(a)\cdot f(b)<0$, a função tem três valores no qual $f(x)=0$.

.

Definição 3.2 (Função monótona)

A $f(x)$ é uma função real contínua monótona num intervalo fechado $[a,b]$ se ela for crescente ou decrescente em valores, tal que

$$f(a)<f(b) \text{ para todo } a<b,$$

ou

$$f(b)>f(a) \text{ para todo } a<b.$$

Esta funções $f(x)$ podem ser chamadas estritamente monótonas para diferenciá-las da que satisfazem

$$f(a)\leq f(b) \text{ para todo } a<b,$$

ou

$$f(b)\geq f(a) \text{ para todo } a<b,$$

que são chamadas fracamente monótonas.

Teorema 3.5 (Unicidade de Raízes)

Se $f(x)$ é uma função contínua e diferenciável em um intervalo num intervalo fechado $[a,b]$ com $f(a)\cdot f(b)<0$ e a derivada $f'(x)$ tem sinal constante $[a,b]$,então existe um único $\xi\in[a,b]$ tal que $f(\xi)=0$.

♡

Exemplo 3.1

f(x) no gráfico abaixo exibe os quatro casos típicos. Observe que $f(a)\cdot f(b)<0$ e $f'(x)\neq 0$.

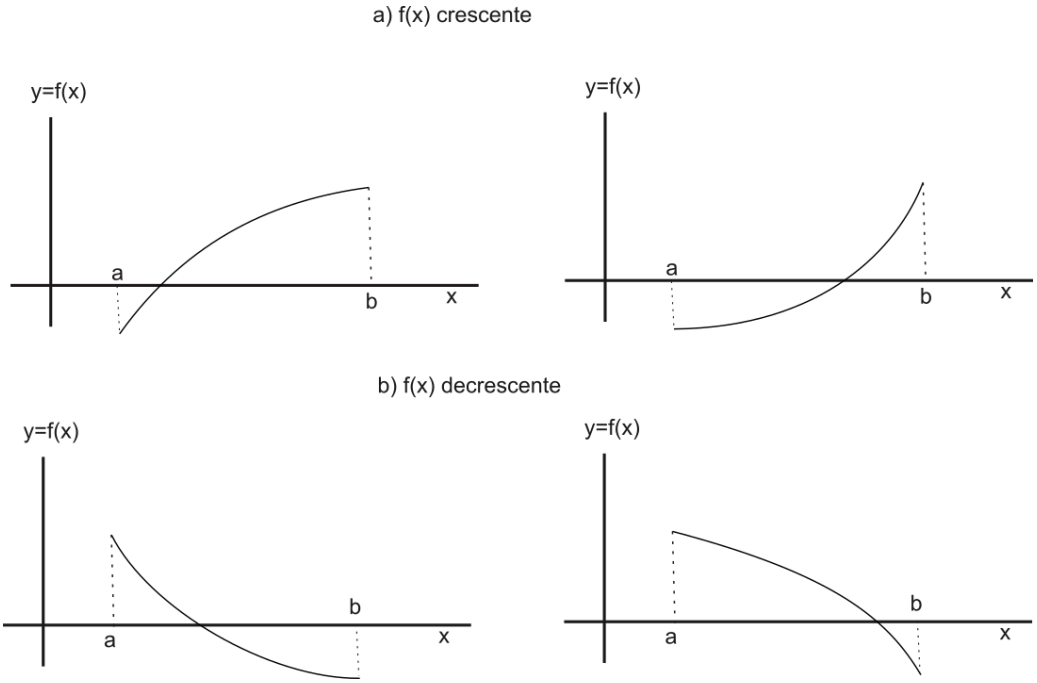


Figura 3.5: Unicidade das Raízes: $f(a)\cdot f(b)<0$ e $f'(x)\neq 0$.

3.2 Solução Numérica de Raizes

A estratégia para obtenção de um raiz real da equação $f(x)=0$ qualquer usando métodos numéricos é realizada em duas etapas:

- Separação ou localização da raiz que correspondem a encontrar $\xi\in[a,b]$ ou uma aproximação inicial $x_0\approx\xi$;
- Utilização de método numérico para encontrar uma raiz com precisão e exatidão aceitável.

3.2.1 Separação/Localização de Raizes

Do teorema da unicidada:
Se $f(x)$ é monótona (crescente ou decrescente, $f'(x)$ não muda de sinal,$f'(x)\neq 0$) em $[a,b]$ e $f(a)\cdot f(b)<0$ então só existe uma raiz ξ da equação $f(x)=0$. Por outro lado, se $f(a)\cdot f(b)>0$ então não existe raiz no intervalo $[a,b]$.

É a situação tipica assumida em aplicações praticas na determinação de raízes.

3.2.2 Localização por Métodos Gráficos

Otenção de uma aroximação inicial $x_0\approx\xi$ ou intervalo onde $\xi\in[a,b]$:

- Traçar Gráficos de $f(x)$ com uso de computadores;
- Utilização de Esboços Gráficos de $g(x)$ e $h(x)$, onde $f(x)=g(x).-h(x)$, ou seja $g(x)=h(x)$ om uso de lápis e papel.

É são as abordagens tipicas em aplicações praticas na determinação de raízes.

3.2.3 Galeria de Esboços Gráficos de Funções

Exemplos de Esboços Gráficos na Localização de Raízes:

1. $f(x) = x^n$, n par.Ver Figura 3.6;
2. $f(x) = x^n$, n impar. Ver Figura 3.7;
3. $f(x) = \dfrac{1}{x^n}$, n impar. Ver Figura 3.8;
4. $f(x) = \dfrac{1}{x^n}$, n par. Ver Figura 3.9;
5. $f(x) = e^x$. Ver Figura 3.10;
6. $f(x) = e^{-x}$. Ver Figura 3.11;
7. $f(x) = ln(x)$. Ver Figura 3.12;
8. $f(x) = log(x)$. Ver Figura 3.13;
9. $f(x) = sin(x)$. Ver Figura 3.14;
10. $f(x) = cos(x)$. Ver Figura 3.15;
11. $f(x) = tg(x)$. Ver Figura 3.16;
12. $f(x) = tgh(x)$, g(x)= senh(x), h(x)=cosh(x). Ver Figura 3.17;

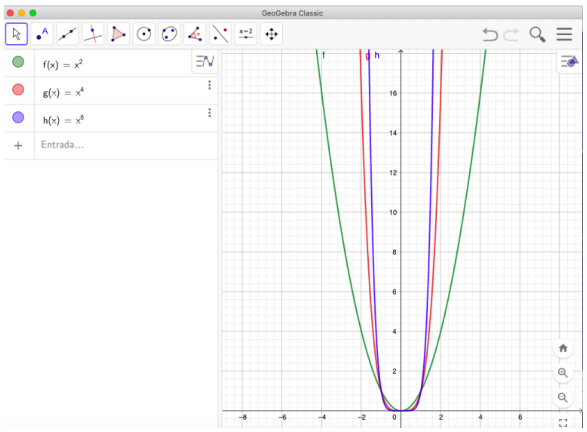


Figura 3.6: $f(x) = x^n$, n par.

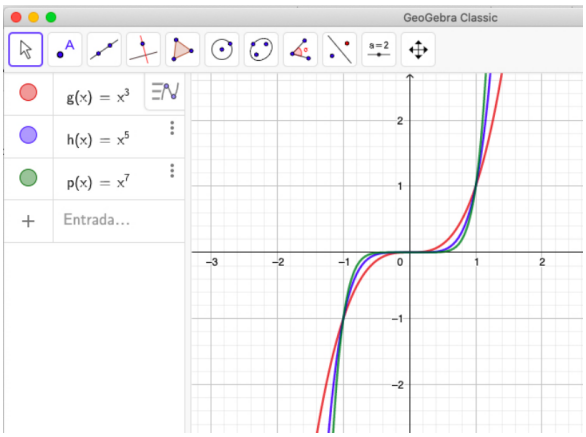


Figura 3.7: $f(x) = x^n$, n impar.

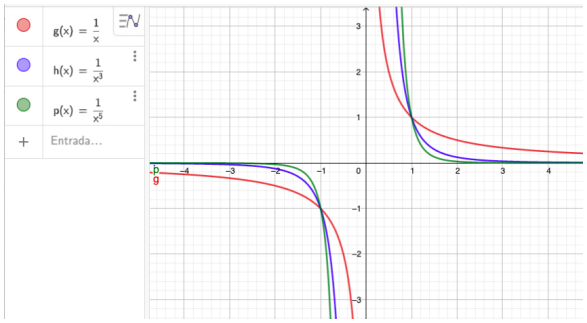


Figura 3.8: $f(x) = \dfrac{1}{x^n}$, n impar.

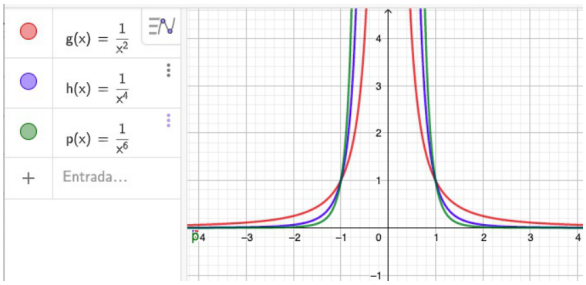


Figura 3.9: $f(x) = \dfrac{1}{x^n}$, n par.

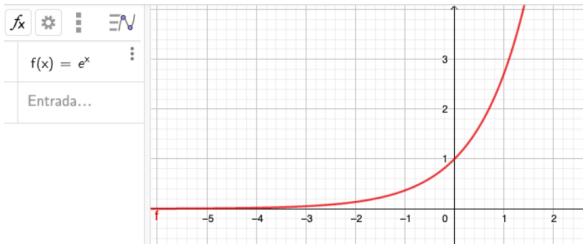


Figura 3.10: $f(x) = e^x$.

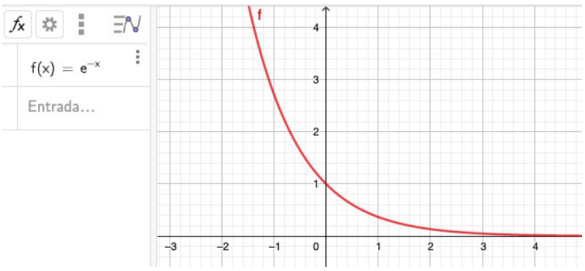


Figura 3.11: $f(x) = e^{-x}$.

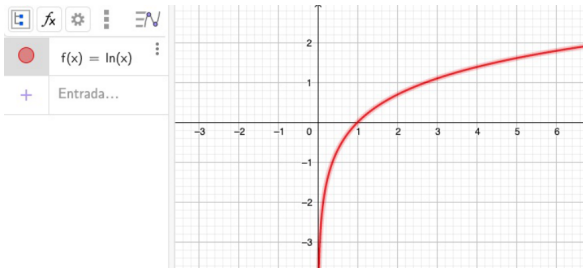


Figura 3.12: $f(x) = ln(x)$.

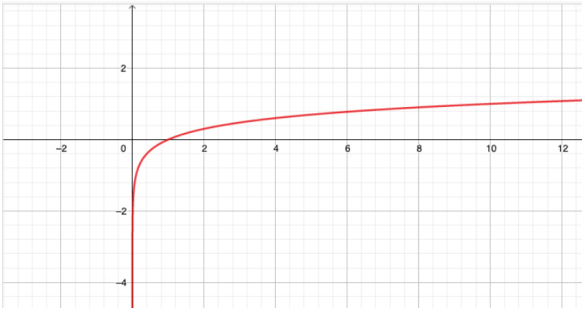


Figura 3.13: $f(x) = log(x)$.

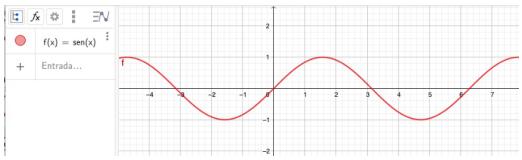


Figura 3.14: $f(x) = sin(x)$.

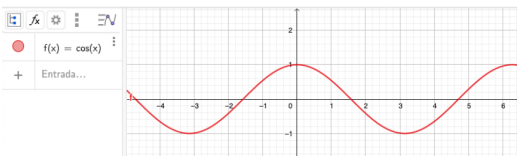


Figura 3.15: $f(x) = cos(x)$.

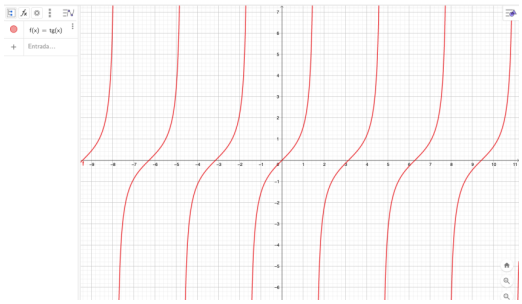


Figura 3.16: $f(x) = tg(x)$.

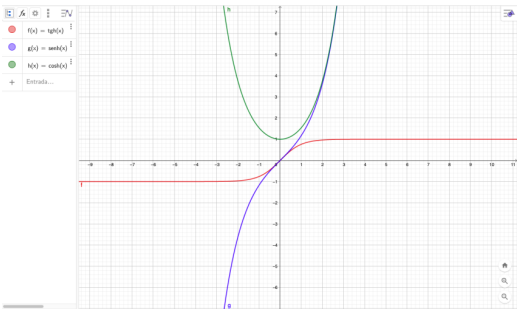


Figura 3.17: $f(x) = tgh(x)$, g(x)= senh(x), h(x)=cosh(x).

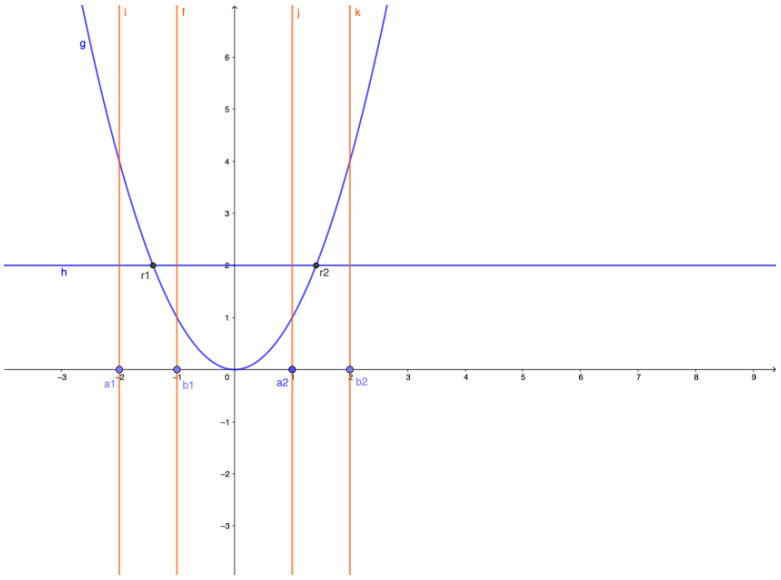
3.2.4 Esboços Gráficos na Localização de Raízes

Exemplos de Esboços Gráficos na Localização de Raízes:

1. $f(x) = x^2 - 2$;
2. $f(x) = x^3 - 2x + 1$;
3. $f(x) = \sqrt{x} - \dfrac{1}{x}$
4. $f(x) = x^2 - cos(x)$

Localização de Raizes: $f(x) = x^2 - 2$

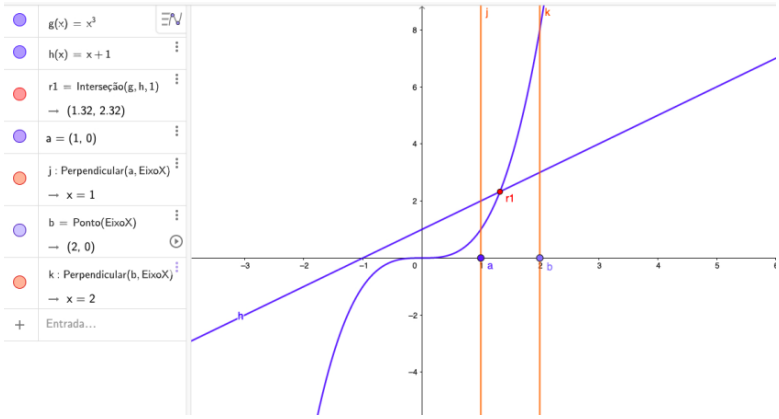
$$f(x) = x^2 - 2 = 0 \therefore x^2 = 2 \therefore g(x) = x^2 \text{ e } h(x) = 2$$



Verificando:
$r_1 \in [-2, -1]$?
$f(-2) = (-2)^2 - 2 = 2$
$f(-1) = (-1)^2 - 2 = -1$
Então $f(-2) \cdot f(-1) < 0$ OK !
$r_2 \in [1, 2]$?
$f(1) = (1)^2 - 2 = -1$
$f(2) = (2)^2 - 2 = 2$
Então $f(1) \cdot f(2) < 0$ OK !

Localização de uma Raiz: $f(x) = x^3 - x - 1 = 0$

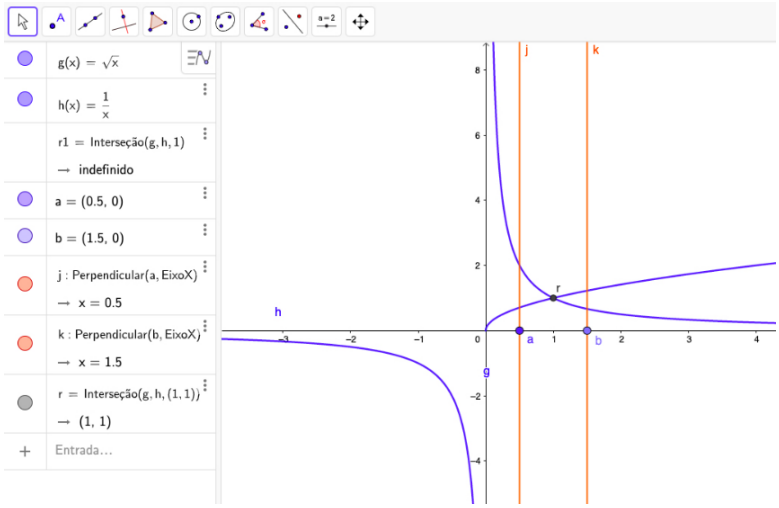
$$f(x) = x^3 - x - 1 = 0 \therefore x^3 = x + 1 \therefore g(x) = x^3 \text{ e } h(x) = x + 1$$



Verificando: $r \in [1, 2]$?
$f(1) = (1)^3 - 1 - 1 = -1$
$f(2) = (2)^3 - 2 - 1 = 5$
Então $f(1) \cdot f(2) < 0$ OK !

Localização de uma Raiz: $f(x) = \sqrt{x} - \dfrac{1}{x} = 0$

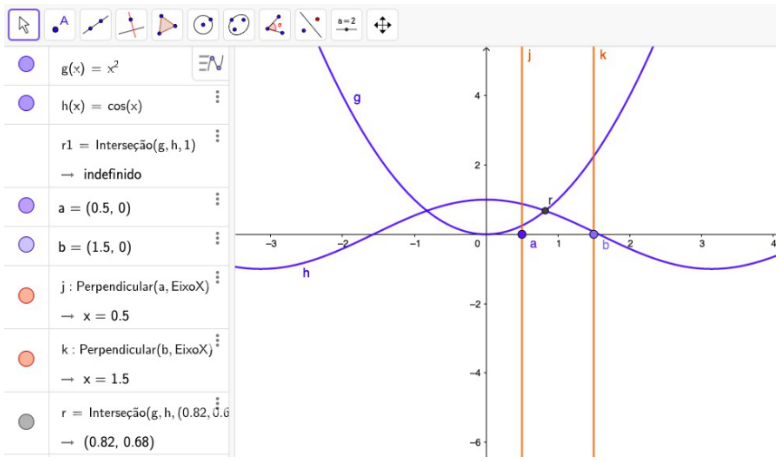
$$f(x) = \sqrt{x} - \frac{1}{x} = 0 \therefore \sqrt{x} = \frac{1}{x} \therefore g(x) = \sqrt{x} \text{ e } h(x) = \frac{1}{x}$$



Verificando: $r \in [0.5, 1.5]$?
$f(0.5) = \sqrt{0.5} - \dfrac{1}{0.5} \approx -1.29$
$f(1.5) = \sqrt{1.5} - \dfrac{1}{1.5} \approx 0.56$
Então $f(0.5) \cdot f(1.5) < 0$ OK !

Localização de uma Raiz: $f(x) = x^2 - cos(x) = 0$

$f(x) = x^2 - cos(x) = 0 \therefore x^2 = cos(x) \therefore g(x) = x^2$ e $h(x) = cos(x)$:



Verificando: $r \in [0.5, 1.5]$?
$f(0.5) = (0.5)^2 - cos(0.5) \approx -0.63$
$f(1.5) = (1.5)^2 - cos(1.5) \approx 2.18$
Então $f(0.5) \cdot f(1.5) < 0$ OK !

3.3 Método da Bisseção

Seja $f(x)$ continua no intervalo $[a, b]$ com $f(a) \cdot f(b) < 0$, então $\exists\, \xi \in [a, b]$. Calcula-se $\bar{x} = (a + b)/2$ e $f(\bar{x})$. Se $f(\bar{x}) = 0$, então $\bar{x}$ é raiz; caso contrário, se $f(\bar{x}) \cdot f(a) < 0$, então $\exists\, \xi \in [a, \bar{x}]$ ou se $f(\bar{x}) \cdot f(b) < 0$, então $\exists\, \xi \in [\bar{x}, b]$. Renomeia-se $\bar{x}$ por a ou b, respectivamente. Existe ξ agora em um intervalo cujo comprimento é a metade do intervalo original. O processo é repetido e para-se a iteração quando $f(\bar{x})$ é muito próximo de zero, ou zero, ou o comprimento do intervalo $[a, b]$ é muito pequeno ϵ, tal que

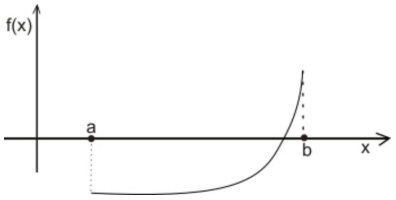
$$|b - a| \leq \epsilon$$

. É fácil de calcular o número de bisseções necessária para alcançar um ϵ prescrito. O intervalo original Δx é reduzido $\Delta x/2$ após um bisseção, $\Delta x/2^2$ depois de duas bisseções, e depois de n bisseções é $\Delta x/2^n$. Colocando $\Delta x/2^n = \epsilon$ e resolvendo para n, tem-se

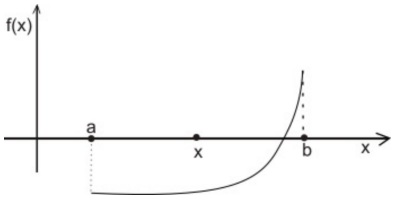
$$n = \frac{ln(|\Delta x| / \epsilon)}{ln(2)} \tag{3.1}$$

Claramente, o método de bisseção converge de forma linear, uma vez que o erro se comporta como $e_{k+1} = e_k/2$.

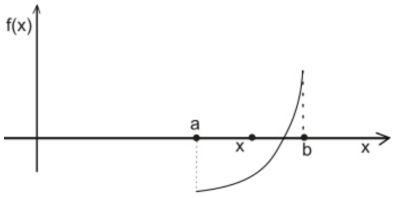
3.3.1 Algoritmo do Método da Bisseção



1 - Comece com uma raiz ξ no intervalo $[a, b]$, *i.e.* $f(a) \cdot f(b) < 0$;



2 - Estime a raiz como o ponto médio do intervalo $x = \frac{(a+b)}{2}$;
3 - Determine o intervalo que contém a raiz. Se $f(x) * f(a) < 0$ então $\xi \in [a, x]$ senão $\xi \in [x, b]$;



4 - Calcule a estimativa do erro;
5 - Repetir os passos $2 \cdots 4$ até que um critério de parada é atingido.

Propiedade

- *Ele sempre converge para uma raiz $\xi \in [a, b]$;*
- *Pegadinha - se $f(a) \cdot f(b) < 0$ e $f(x)$ tem mais de uma raiz em $[a, b]$.*

Exemplo 3.2

Encontre a raiz cúbica de 3.

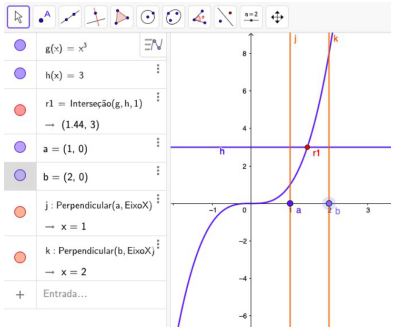
$x^3 = 3 \therefore f(x) = x^3 - 3.$

A solução é 1.442249570307408 obtida com a calculadora do celular !

Vamos apresentar os detalhes dos cálculos, em duas etapas, para localização da raiz e aplicação do Método da Bisseção.

1^a *Etapa: Separação da Raiz.*

$x^3 - 3 \therefore x^3 = 3 \therefore g(x) = x^3$ e $h(x) = 3$:



Verificando:
$r \in [1, 2]$?
$f(1) = (1)^3 - 3 \approx -2$
$f(2) = (2)^3 - 3 \approx 5$
Então $f(1) \cdot f(2) < 0$ OK !

2^a Etapa: Estimativa de $\xi \in [1, 2]$ usando o Método da Bisseção.

$f(1) = -1(-),\ f(2) = 5(+)$

- *k=1*, $\xi \in [\overset{(-)}{1}, \overset{(+)}{2}]$
  $x = \dfrac{1+2}{2} = 1.5$
  $f(1.5) = (1.5)^3 - 3 = 0.38(+)$
- *k=2*, $\xi \in [\overset{(-)}{1}, \overset{(+)}{1.5}]$
  $x = \dfrac{1+1.5}{2} = 1.25$
  $f(1.25) = (1.25)^3 - 3 = -1.05(-)$
- *k=3*, $\xi \in [\overset{(-)}{1.25}, \overset{(+)}{1.5}]$
  $x = \dfrac{1.25+1.5}{2} = 1.38$
  $f(1.38) = (1.38)^3 - 3 = -0.37(-)$
- *k=4*, $\xi \in [\overset{(-)}{1.38}, \overset{(+)}{1.5}]$
  $x = \dfrac{1.38+1.5}{2} = 1.44$
  $f(1.44) = (1.44)^3 - 3 = -0.014(-)$
- *k=5*, $\xi \in [\overset{(-)}{1.44}, \overset{(+)}{1.5}]$
  $x = \dfrac{1.44+1.5}{2} = 1.47$
  $f(1.47) = (1.47)^3 - 3 = 0.17(+)$
- *k=6*, $\xi \in [\overset{(-)}{1.44}, \overset{(+)}{1.5}]$
  $x = \dfrac{1.44+1.47}{2} = 1.46$
  $f(1.46) = (1.46)^3 - 3 = 0.11(+)$

Tabela 3.1: Método da Bisseção (Resumo).

k	$x_{[k]}$	$f(x_{[k]})$	$Erro_{[k]}$
1	1.5	0.38	–
2	1.25	-1.05	0.25
3	1.38	-0.37	0.13
4	1.44	-0.014	0.06
5	1.47	0.17	0.03
6	1.46	0.11	0.01

Definição: $Erro_{[k]} = \left|x_{[k]} - x_{[k-1]}\right|$

3.3.2 Convergência

Considerando $a_0 = a$, $b_0 = b$ e $[a_n, b_n](n = 0,\ 1,\ 2,\ \cdots)$ os intervalos sucessivos das bisseções tem-se que

$$a_0 \leq a_1 \leq a_2 \leq a_3 \leq \cdots \leq b_0 = b$$

e

$$b_0 \geq b_1 \geq b_2 \geq b_3 \geq \cdots \geq a_0 = a$$

A sequencia $\{a_n\}$ é monótona crescente e limitada acima e a sequencia $\{b_n\}$ é monótona decrescente e limitada abaixo. Portanto as duas sequencias convergem. Como

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2} = \cdots = \frac{b-a}{2^n} \tag{3.2}$$

Calculando o limite, a equação (3.2) tem-se

$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{b-a}{2^n} = (b-a) \lim_{n\to\infty} \frac{1}{2^n} = 0 \tag{3.3}$$

Conclui então que as sequencias $\{a_n\}$ e $\{b_n\}$ têm o mesmo limite

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \xi \tag{3.4}$$

Agora falta mostrar que ξ é uma raiz da função $f(x)$. Do algoritmo da bisseção tem-se que $f(a_n) \cdot f(b_n) \leq 0$. Então tomando o limite

$$\lim_{n\to\infty} f(a_n) \cdot f(b_n) \leq 0$$

Finalmente, usando (3.4), tem-se

$$[f(\xi))]^2 \leq 0 \quad \Longrightarrow \quad f(\xi) = 0$$

Em resumo, o método da bisseção sempre converge para raiz des que o intervalo inicial contenha raiz, e fornece a raiz de f(x).

3.3.3 Ordem de convergência

$$\text{De } |e_{n+1}| = |\xi - \bar{x}_{n+1}| \leq \frac{1}{2}(b_{n+1} - a_{n+1}) = \frac{1}{2}\frac{(b_n - a_n)}{2} \quad \text{e} \quad |e_n| = |\xi - \bar{x}_n| \leq \frac{1}{2}(b_n - a_n)$$

(Notar que este limitante é independente da função f(x) e/ou de suas derivadas.) Portanto tem-se

$$|e_{n+1}| \cong \frac{1}{2}|e_n|$$

Conclui-se que o método da bisseção tem convergência linear.

3.4 Regula Falsi

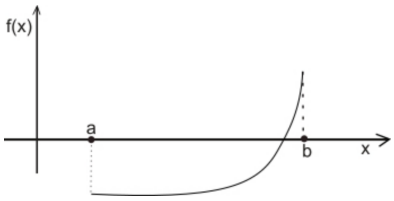
O método da *regula falsi* é semelhante ao da Bisseção cuja diferença é o cálculo da aproximação $\bar{x}$ no intervalo $[a, b]$ no qual tem-se $f(\xi) = 0$, onde f(x) contínua com $f(a)$ e $f(b)$ tendo sinais opostos. Aproximando-se $f(x)$ em $[a, b$ por uma reta passando pelos pontos $(a, f(a))$ e $b, f(b)$, então a solução aproximada é calculada como o ponto o a reta cruza a eixo dos x:

$$\bar{x} = a - \frac{f(a)}{f(b) - f(a)}(b - a) = b - \frac{f(b)}{f(b) - f(a)}(b - a) = \frac{af(b) - bf(a)}{f(b) - f(a)} \tag{3.5}$$

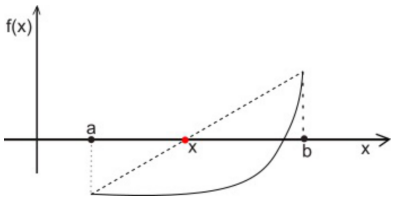
A modificação acima pode ser interpretado como uma média ponderada. Assumindo-se $f(a) \cdot f(b) < 0$ então (3.5) pode ser escrito como

$$\bar{x} = \frac{|f(b)|}{|f(b)| + |f(a)|}a + \frac{|f(a)|}{|f(b)| + |f(a)|}b$$

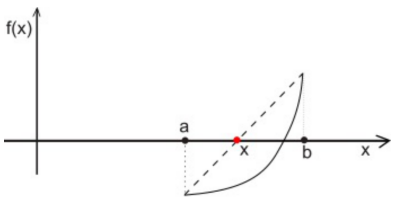
3.4.1 Algoritmo do Método da Regula Falsi



1 - Comece com uma raiz ξ no intervalo $[a, b]$, *i.e.* $f(a) \cdot f(b) < 0$;



2 - Estime a raiz no intervalo como: $x = \frac{a \cdot f(b) - b \cdot f(a)}{f(b) - f(a)}$;
3 - Determine o intervalo que contém a raiz. Se $f(x) * f(a) < 0$ então $\xi \in [a, x]$ senão $\xi \in [x, b]$;



4 - Calcule a estimativa do erro;
5 - Repetir os passos $2 \cdots 4$ até que um critério de parada é atingido.

Propiedade

- $x = \dfrac{a \cdot f(b) - b \cdot f(a)}{f(b) - f(a)}$, *com* $f(a) \cdot f(b) < 0$, *então*
  $x = \dfrac{a \cdot |f(b)| + b \cdot |f(a)|}{|f(b)| + |f(a)|}$ *é uma média ponderada;*
- *Pegadinha - se a estimativa* $x \notin [a, b]$ *é que algo está errado no cálculo.*

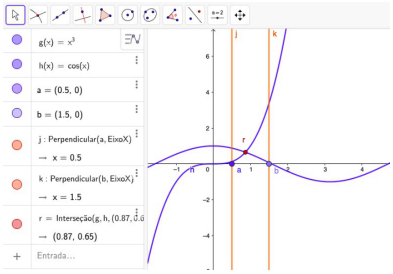
Exemplo 3.3

Encontre uma raiz de $x^3 - cos(x)$.

Vamos apresentar os detalhes dos cálculos, em duas etapas, para localização da raiz e aplicação do Método da *Regula Falsi* (também conhecido como Método das Cordas ou Método da Posição Falsa).

1^a *Etapa: Separação da Raiz.*

$x^3 - cos(x) \therefore x^3 = cos(x) \therefore g(x) = x^3$ e $h(x) = cos(x)$:



Verificando: $r \in [0.5, 1.5]$?
$f(0.5) = (0.5)^3 - cos(0.5) \approx -0.75$
$f(1.5) = (1.5)^3 - cos(1.5) \approx 3.30$
Então $f(0.5) \cdot f(1.5) < 0$ OK !

Obs: A máquina deve estar em radianos. O gráfico foi feito em radianos.

2^a *Etapa:*
Estimativa de $\xi \in [0.5, 1.5]$ *usando o Método da Regula Falsi .*

$f(0.5) = -0.75(-),\ f(1.5) = 3.30(+)$

- *k=1*, $\xi \in \overset{(-)}{[0.5}, \overset{(+)}{1.5}]$
  $x = \dfrac{0.5 \cdot f(1.5) - 1.5 \cdot f(0.5)}{f(1.5) - f(0.5)} = 0.68$
  $f(0.68) = (0.68)^3 - cos(0.68) = -0.46(-)$
- *k=2*, $\xi \in \overset{(-)}{[0.68}, \overset{(+)}{1.5}]$
  $x = \dfrac{0.68 \cdot f(1.5) - 1.5 \cdot f(0.68)}{f(1.5) - f(0.68)} = 0.78$
  $f(0.78) = (0.78)^3 - cos(0.78) = -0.24(-)$
- *k=3*, $\xi \in \overset{(-)}{[0.78}, \overset{(+)}{1.5}]$
  $x = \dfrac{0.78 \cdot f(1.5) - 1.5 \cdot f(0.78)}{f(1.5) - f(0.68)} = 0.83$
  $f(0.83) = (0.83)^3 - cos(0.83) = -0.10(-)$
- *k=4*, $\xi \in \overset{(-)}{[0.83}, \overset{(+)}{1.5}]$
  $x = \dfrac{0.83 \cdot f(1.5) - 1.5 \cdot f(0.83)}{f(1.5) - f(0.83)} = 0.85$

$f(0.85) = (0.85)^3 - cos(0.85) = -0.045(-)$

- *k=5*, $\xi \in [\overset{(-)}{0.85}, \overset{(+)}{1.5}]$

$$x = \frac{0.85 \cdot f(1.5) - 1.5 \cdot f(0.85)}{f(1.5) - f(0.85)} = 0.86$$

$f(0.86) = (0.86)^3 - cos(0.86) = -0.016(-)$

Tabela 3.2: Método da *Regula Falsi* (Resumo).

k	$x_{[k]}$	$f(x_{[k]})$	$Erro_{[k]}$
1	0.68	-0.46	–
2	0.78	-0.24	0.10
3	0.83	-0.10	0.05
4	0.85	-0.045	0.02
5	0.86	-0.016	0.01

Definição: $Erro_{[k]} = \left|x_{[k]} - x_{[k-1]}\right|$

3.4.2 Convergência

Sejam $[a_k, b_k]$ os intervalos sucessivos do método da *regula falsi*. Geralmente $b_k - a_k$ não tende para zero com $k \to \infty$ e mesmo assim o método converge para a raiz $\xi \in [a, b]$. O método da *regula falsi* converge linearmente, quase sempre com uma convergência unilateral.

3.5 Método das Secantes

O método da secante é um procedimento de busca de raízes na análise numérica que usa uma série de raízes de linhas secantes para melhor aproximar uma raiz de uma função f. Vamos aprender mais sobre o método da secante, sua fórmula, vantagens e limitações, e exemplo resolvido do método da secante.

3.5.1 O que é o Método da Secante?

A linha tangente à curva de $y = f(x)$ com o ponto de tangência $(x_0, f(x_0))$ é utilizada na abordagem de Newton. O gráfico da linha tangente em torno de $x = \alpha$ é essencialmente o mesmo que o gráfico de $y = f(x)$ quando $x_0 \approxeq \alpha$. A raiz da linha tangente foi usada para aproximar a raiz α . Considere empregar uma linha de aproximação baseada na 'interpolação'. Vamos supor que temos duas estimativas de raiz para ξ, digamos, x_0 e x_1. Então, temos uma função linear

$$q(x) = a_0 + a_1 x$$

usando $q(x_0) = f(x_0), q(x_1) = f(x_1)$. Essa linha também é conhecida como linha secante. Sua fórmula é a seguinte:

$$q(x) = \frac{(x_1 - x)f(x_0) + (x - x_0)f(x_1)}{x_1 - x_0}$$

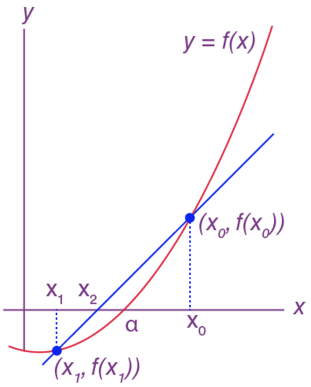
A equação linear $q(x) = 0$ é resolvida, com a raiz denotada por x_2:

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

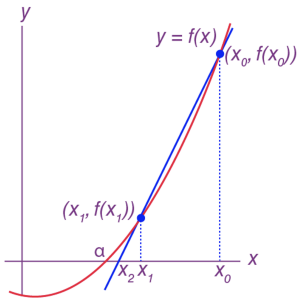
ou

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

3.5.2 Algoritmo do Método das Secantes



$f(x_0) > 0, f(x_1) < 0$



$f(x_0) > 0, f(x_1) > 0$

1- Comece com x_0 e x_1, valores próximos a uma raiz ξ, tomados como aproximações iniciais (veja exemplos gráficos à esquerda);

2- Iteração: $(k = 1, 2, 3, \dots)$ Estime a raiz no intervalo como: $x_{k+1} = \frac{x_{k-1} \cdot f(x_k). - x_k \cdot f(x_{k-1})}{f(x_k) - f(x_{k-1})}$
até que um critério de parada seja atingido (ou seja, a precisão desejada da resposta ou o número máximo de iterações tenha sido alcançado).

3.5.3 Convergência do Método das Secantes

- Se os valores iniciais x_0 e x1 estiverem próximos o suficiente da raiz, o método da secante itera xn e converge para uma raiz da função f. A ordem de convergência é dada por φ, onde;

$$\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618,$$

  é a razão áurea.
- A convergência é particularmente superlinear, mas não é quadrática. Esta solução é válida apenas sob certos requisitos técnicos, como $f(x)$ ser duas vezes continuamente diferenciável e a raiz ser simples na questão (ou seja, ter multiplicidade 1).
- Não há garantia de que o método da secante convergirá se os valores iniciais não estiverem próximos o suficiente da raiz. Por exemplo, se a função $f(x)$ for diferenciável no intervalo $[x_0, x_1]$, e houver um ponto no intervalo onde $f'(x) = 0$, o algoritmo pode não convergir.

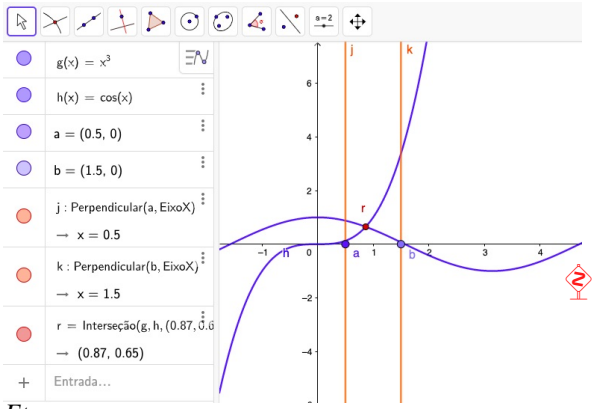
Exemplo 3.4

Encontre uma raiz de $x^3 - cos(x)$.

Vamos apresentar os detalhes dos cálculos, em duas etapas:
1) determinação de x_0 e x_1 tomados como aproximações iniciais para da raiz;
2) aplicação do Método das Secantes (também conhecido como Método dos 2 Pontos).

1^a *Etapa:*

Escolha de x_0 *e* x_1*, valores próximos a uma raiz* . $x^3 - cos(x) = 0 \therefore x^3 = cos(x) \therefore g(x) = x^3$ e $h(x) = cos(x)$:



Escolhendo:

$x_0 = 1.5$

$x_1 = 1.0$

Valores próximos a raiz ξ.

Nota

A máquina deve estar em radianos. O gráfico foi feito em radianos.

2^a *Etapa:*

$x_0 = 1.5, \quad f(1.5) = 3.30$

$x_1 = 1.0, \quad f(1.0) = 0.46$

- *Iteração =1*

$$x_2 = \frac{1.5 \cdot f(1.0) - 1.0 \cdot f(1.5)}{f(1.0) - f(1.5)}$$

$$x_2 = \frac{1.5 \cdot 0.46 - 1.0 \cdot 3.30}{0.46 - 3.30} = 0.92$$

$$f(0.92) = (0.92)^3 - cos(0.92) = 0.17$$

- *Iteração=2*

$$x_3 = \frac{1.0 \cdot f(0.92) - 0.92 \cdot f(1.0)}{f(0.92) - f(1.0)}$$

$$x_3 = \frac{1.0 \cdot 0.17 - 0.92 \cdot 0.46}{0.17 - 0.46} = 0.87$$

$$f(0.87) = (0.87)^3 - cos(0.87) = 0.014$$

- *Iteração=3*

$$x_4 = \frac{0.92 \cdot f(0.87) - 0.87 \cdot f(0.92)}{f(0.87) - f(0.92)}$$

$$x_4 = \frac{0.92 \cdot 0.014 - 0.87 \cdot 0.17}{0.014 - 0.17} = 0.866$$

$$f(0.866) = (0.866)^3 - cos(0.866) = 1.58 \cdot 10^{-3}$$

- *Iteração=4*

$$x_5 = \frac{0.87 \cdot f(0.866) - 0.866 \cdot f(0.87)}{f(0.866) - f(0.87)}$$

$$x_5 = \frac{0.87 \cdot 1.58 \cdot 10^{-3} - 0.866 \cdot 0.014}{1.58 \cdot 10^{-3} - 0.014} = 0.8655$$

$$f(0.8655) = (0.8655)^3 - cos(0.8655) = 7.8124 \cdot 10^{-5}$$

Tabela 3.3:
Método das Secantes (Resumo)

k	$x_{[k]}$	$f(x_{[k]})$	$Erro_{[k]}$
0	1.5	3.30	–
1	1.0	0.46	–
2	0.92	0.17	0.08
3	0.87	0.014	0.5
4	0.866	$1.58 \cdot 10^{-3}$	0.004
5	0.8655	$7.81 \cdot 10^{-5}$	0.0005

Definição:

$$Erro_{[k]} = \left| x_{[k]} - x_{[k-1]} \right|$$

3.5.4 Vantagens e Desvantagens do Método da Secante

O método da secante possui as seguintes vantagens:

- Converge mais rapidamente que uma taxa linear, tornando-o mais convergente do que o método da bisseção.
- Não necessita do uso da derivada da função, que pode não estar disponível em várias aplicações.
- Ao contrário da técnica de Newton, que requer duas avaliações da função em cada iteração, requer apenas uma.

O método da secante possui as seguintes desvantagens:

- O método da secante pode não convergir.
- As iterações calculadas não têm garantias de limites de erro.
- Se $f'(\alpha) = 0$, é provável que seja desafiador. Isso significa que, quando $x = \alpha$, o eixo x é tangente ao gráfico de $y = f(x)$.
- A abordagem de Newton é mais facilmente generalizada para novas formas de resolver sistemas de equações simultâneas não lineares.

3.6 Método de Newton-Raphson

Seja x_0 um valor inicial próximo a raiz ξ de $f(x) = 0$. Considere h a correção $\xi = x_0 + h$. Então $f(\xi) = 0$ implica que $f(x_0 + h) = 0$. Seja h pequeno e $f(x)$ tenha segunda derivada continua, então

$$f(\xi) = f(x_0) + hf^{'}(x_0) + \frac{h^2}{2}f^{'}(\eta) = 0, \quad \eta \in [x_0, x_0 + h]$$

Desprezando os termos quadráticos e de ordem superiores e assumindo que ξ é uma raiz simples, encontra-se

$$h \approx -\frac{f(x_0)}{f^{'}(x_0)} \Rightarrow x_1 = x_0 - \frac{f(x_0)}{f^{'}(x_0)}$$

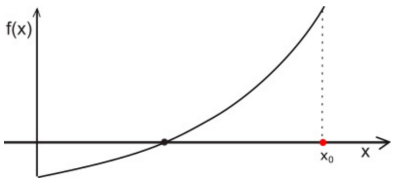
deve ser uma melhor aproximação para ξ que x_0. Continuando este processo com $x_1, x_2 \cdots$ o método é dado por

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f^{'}(x_{k-1})}, \quad k = 1, 2, \cdots n$$

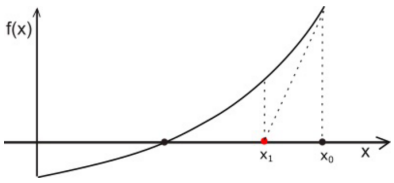
Geometricamente, x_k é a interseção da tangente com o eixo-x que passa pelo ponto $(x_{k-1},\ f(x_{k-1}))$. Este método pode não convergir para a raiz desejada se o valor inicial é muito distante da raiz.

3.6.1 Algoritmo do de Newton-Raphson

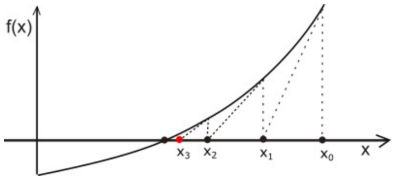
1 - Comece com uma aproximação x_0 para a raiz ξ;



2 - Estime a raiz no intervalo como $x_1 = x_0 - \dfrac{f(x_0)}{f'(x_0)}$;
3 - Calcule a estimativa do erro;



4 - Repetir para $k = 2 \cdots n$
a) $x_k = x_{k-1} - \dfrac{f(x_{k-1})}{f'(x_{k-1})}$;
b) calcule a estimativa do erro
até que um critério de parada é atingido.



Propriedades:

- Tem convergência quadrática para raízes simples;
- Em geral, o método de Newton-Raphson tem um bom desempenho, com uma boa escolha da aproximação inicial x_0;
- A escolha inadequada da aproximação inicial x_0 pode gerar uma sequencia que converge para uma raiz diferente da procurada;
- Requer a avaliação da função e sua derivada- tendo um custo computacional adicional e exigindo que a derivada da função exista;
- Pegadinha: Apesar de sua popularidade, o método newton não é universalmente elogiado. No livro de Hamming tem uma seção com o título ”Método de Newton (outro método a evitar)- ”Newton method (another method to avoid);.

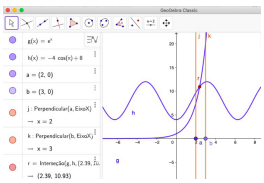
Exemplo 3.5

Encontre uma raiz de $e^x + 4cos(x) - 8 = 0$.

Será apresentado os detalhes dos cálculos, em duas etapas, para localização da raiz e aplicação do Método de Newton.

1^a **Etapa: Separação da Raiz.**

$$e^x + 4cos(x) - 8 \therefore e^x = -4cos(x) + 8 \therefore g(x) = e^x \text{ e } h(x) = -4cos(x) + 8 :$$



Verificando: $r \in [2, 3]$?
$f(2) = e^2 + 4cos(2) - 8 \approx -2.27$
$f(3) = e^3 + 4cos(3) - 8 \approx 8.12$
Então $f(2) \cdot f(3) < 0$ OK !

A aproximação inicial escolhida foi $x_0 = 2.4$, valor próximo a raiz.

Obs: A máquina deve estar em radianos. O gráfico foi feito em com cálculo do $cos(x)$ em radianos.

2^a **Etapa: Determinação de ξ usando o Método de Newton**.

Repetir para $k = 1 \cdots n$
$x_k = x_{k-1} - \dfrac{f(x_{k-1})}{f'(x_{k-1})}$; $x_0 = 2.4$;
$f(x) = e^x + 4cos(x) - 8$; $f'(x) = e^x - 4sen(x)$

- **k=1**
  $x_1 = 2.4 - \dfrac{f(2.4)}{f'(2.4)} = 2.39$
  $f(2.4) = e^{2.4} + 4cos(2.4) - 8 = 0.0736$
  $f'(2.4) = e^{2.4} - 4sen(2.4) = 8.32$

- **k=2**
  $x_2 = 2.39 - \dfrac{f(2.39)}{f'(2.39)} = 2.3900$
  $f(2.39) = e^{2.39} + 4cos(2.39) - 8 = -0.008915$
  $f'(2.39) = e^{2.39} - 4sen(2.39) = 8.1823$

- **k=3**
  $x_3 = 2.3900 - \dfrac{f(2.3900)}{f'(2.3900)} = 2.3911$
  $f(2.3900) = e^{2.3900} + 4cos(2.3900) - 8 = -0.00891536$
  $f'(2.3900) = e^{2.3900} - 4sen(2.3900) = 8.182281$

Tabela 3.4: Método de Newton (Resumo).

k	$x_{[k]}$	$f(x_{[k]})$	$Erro_{[k]}$
k	$x_{[k]}$	$f(x_{[k]})$	$Erro_{[k]}$
1	2.39	-0.008915	–
2	2.3900	-0.00891536	0.01
3	2.3911	0.00009351821	0.0011

Definição: $Erro_{[k]} = \left|x_{[k]} - x_{[k-1]}\right|$

3.6.2 Convergência

Dada uma função $f(x)$, se $f(\xi) = 0$ e também é verdade que $f'(\xi) = ... = f^{(m-1)}(\xi) = 0$, mas $f^{(m)}(\xi) \neq 0$, então ξ é dito ser uma raiz de $f(x)$ da ordem de multiplicidade m.

Se ξ for uma raiz simples de $f(x) = 0$, então o método de Newton converge quadraticamente, de modo que cada iteração quase duplica o número de casas decimais exatas. Se, pelo contrário, ξ for uma raiz múltipla, o erro em cada etapa é uma fração do erro na etapa anterior. Ou seja, nesses casos, o método converge linearmente.

Capítulo Execícios

1. Localizar pelo menos uma raiz real das equações abaixo usado o Geogebra: https://www.geogebra.org/
   (a) $x^3 - 6x^2 - x - 30 = 0$
   (b) $x + log(x) = 0$
   (c) $x + 2cos(x) = 0$
   (d) $x^2 - 10 \cdot ln(x) - 5 = 0$
   (e) $x^3 - e^{2x} + 3 = 0$
   (f) $2x^3 + x^2 - 2 = 0$
   (g) $sen(x) - lx(x) = 0$
   (h) $e^{cos(x)} + x^3 - x = 0$
   (i) $2 \cdot ln(3 - cos(x)) - 3x^x + 5 \cdot sen(x) = 0$
   (j) $cos(x) + ln(x) + x = 0$

2. Dado a equação: $3x - cos(x) = 0$
   (a) Localizar a primeira raiz (positiva ou negativa);
   (b) Aplique 5 iterações do um método da Bisseção para encontrar o valor aproximado da raiz.

   Atenção:
   (i) usar 02 casas decimais e no mínimo com 02 dígitos significativos;
   (ii) exibir cálculos detalhados dos valores de $x^{[k]}$, $f^{[k]}$ e $Erro^{[k]}$ encontrados nas iterações;
   (iii) faça tabela resumo exibindo $[k]$, $x^{[k]}$, $f^{[k]}$ e $Erro^{[k]}$.

3. Dado a equação: $e^{3 \cdot x} \cdot (x - 1) + x^3 = 0$
   (a) Verificar que $\xi \in [0.7,\ 1.3]$;
   (b) Aplique 5 iterações do um método da Regula-Falsi para encontrar o valor aproximado da raiz.

   Atenção:
   (i) usar 02 casas decimais e no mínimo com 02 dígitos significativos;
   (ii) exibir cálculos detalhados dos valores de $x^{[k]}$, $f^{[k]}$ e $Erro^{[k]}$ encontrados nas iterações;
   (iii) faça tabela resumo exibindo $[k]$, $x^{[k]}$, $f^{[k]}$ e $Erro^{[k]}$.

4. Dado a equação: $x^2 + sin(x/5) - 1/5$
   (a) Utilizar método gráfico para determinar uma aproximação x_0 de uma raiz ξ ;
   (b) Aplique 3 iterações do um método de Newton para encontrar o valor aproximado da raiz.

   Atenção:
   (i) exibir cálculos detalhados dos valores de $x^{[k]}$, $f(x)^{[k]}$,$f'(x)^{[k]}$, e $Erro^{[k]}$ encontrados nas iterações;
   (ii) faça tabela resumo exibindo $[k]$, $x^{[k]}$, $f^{[k]}$ e $Erro^{[k]}$.

Capítulo - Interpolação

4.1 Introdução

Sempre que se procura numa tabela por valor que não consta explicitamente na tabela, efetua-se uma interpolação, quando o valor está entre o menor e o maior valor presente na tabela, para a valor pesquisado. Se o valor estiver fora do intervalo presente na tabela, tem-se então uma problema de extrapolação.

Denomina-se interpolação um método que permite construir um novo conjunto de dados a partir de um conjunto discreto de dados pontuais previamente conhecidos. Em engenharia e ciências, dispõe-se habitualmente de conjuntos de dados pontuais, obtidos a partir de uma amostragem ou experimento. Através da interpolação pode-se construir uma função que aproximadamente se "ajuste" nestes dados pontuais. Os conjunto de dados pontuais não possui continuidade, e isto muitas vezes impossibilita a representação teórica de um fenômeno real empiricamente observado.

Usando-se interpolação, pode-se construir uma função que represente estes dados pontuais, conferindo-lhes, então, a continuidade desejada.

Uma aplicação da interpolação é a aproximação de funções complexas por funções mais simples. Obviamente, quando se utiliza uma função mais simples para calcular novos dados, normalmente não se obtém o mesmo resultado da função original, mas dependendo do domínio do problema e do método de interpolação utilizado, o erro é aceitável e tem-se o ganho de simplicidade.

A interpolação permite fazer a computação aproximada de valores de uma função, bastando para tanto conhecer apenas algumas das suas abscissas e respectivas ordenadas (imagens no contra-domínio da função). A função resultante garantidamente passa pelos pontos fornecidos, e, em relação aos outros pontos, pode ser considerada um ajuste.

A aproximação de funções por polinômios é muita usada em métodos numéricos. Isso acontece porque os polinômios são facilmente computáveis, suas derivadas e integrais são também polinômios, suas raízes podem ser determinadas com relativa facilidade, etc.

A aproximação polinomial pode ser obtida de várias maneiras, entre os quais pode-se citar: Interpolação, Método dos Mínimos Quadrados, Osculação, Mini-Max, etc, portanto pode ser proveitoso substituir uma função complicada por um polinômio que a represente. ressalta-se que o Teorema de Weirstrass que afirma que: toda função contínua pode ser aproximada por um polinômio. Neste capitulo será visto como aproximar uma função usando Métodos de Interpolação Polinomial.

4.2 Preliminares

Vamos introduzir a interpolação a partir da tabela:

Pares ordenados (x,y)

x	y
x_0	y_0
x_1	y_1
$\vdots$	$\vdots$
x_n	y_n

$x_0 < x_1 < \cdots < x_n$, e $x \in [x_0, x_n]$
$y = f(x)$, x é o valor a ser interpolado
$f(x) \approx \phi(x)$, onde $y_i = \phi(x_i)$ é a função interpolante, onde só iremos estudar a *Interpolação Polinomial* na variável x.

4.2.1 Aplicações da Interpolação

Para métodos numéricos interpolação é usada para:

- Interpolar valores de tabelas (valores não tabulados);
- Desenvolver fórmulas de integração numérica;
- Desenvolver métodos de diferenciação numérica;
- Desenvolver métodos de para determinação de raízes (interpolação inversa/direta);
- Desenvolver métodos de Elementos Finitos.

Interpolação não é normalmente usada para descrição funcional de dados experimentais pois os erros nos dados podem gerar uma representação inadequada:

- *Ajuste de Curvas* e o tratamento estatístico são mais adequados para dados experimentais.

4.2.2 Outras formas de Interpolação

- Interpolação racional é a interpolação por funções racionais usando o aproximante Padé
- Interpolação trigonométrica é a interpolação por polinômios trigonométricos usando a série de Fourier. Outra possibilidade é usar *wavelets*.
- Interpolação de Whittaker-Shannon pode ser usada se o número de pontos de dados for infinito ou se a função a ser interpolada tiver suporte compacto.

4.2.3 Interpolação de Função de Base Radial

A interpolação de função de base radial (RBF) é um método avançado na teoria de aproximação para a construção de interpolantes precisos de alta ordem de dados não estruturados, possivelmente em espaços de alta dimensão. O interpolante assume a forma de uma soma ponderada de funções de base radial. A interpolação RBF é um método sem malha, o que significa que os nós (pontos no domínio) não precisam estar em uma grade estruturada e não requer a formação de uma malha. Frequentemente, é espectralmente preciso e estável para um grande número de nós, mesmo em grandes dimensões.

A interpolação RBF é usada para aproximar operadores diferenciais, operadores integrais e operadores diferenciais de superfície. Esses algoritmos foram usados em soluções altamente precisas de muitas equações diferenciais, incluindo as equações de Navier-Stokes, a equação de Cahn-Hilliard e as equações de águas rasas.

https://en.wikipedia.org/wiki/Radial_basis_function_interpolation

4.3 Interpolação Linear

Denomina-se interpolação linear o método de interpolação que se utiliza de uma função linear $y(x)$ (um polinômio de primeiro grau) para representar, por aproximação, uma função f(x) (ou suposta f(x) que representa um fenômeno real empiricamente observado) contido no domínio de f(x). A interpolação linear é uma linha que se ajusta a dois pontos. A interpolação linear é mostrada na Figura 1 onde y_0 e y_1 são os valores conhecidos de $y(x) = f(x)$ em $x = x_0$ e $x = x_1$, respectivamente.

O modelo linear é dado por

$$y(x) = a \cdot x + b \tag{4.1}$$

(um polinômio de primeiro grau) Do par de pontos (x_0, y_0) e (x_1, y_1) substituídos na equação 4.1 tem-se

$$\begin{cases} a \cdot x_0 + b = y_0 \\ a \cdot x_1 + b = y_1 \end{cases}$$

Acrescentado a equação 4.1 obtém-se

$$\begin{cases} a \cdot x_0 + b = y_0 \\ a \cdot x_1 + b = y_1 \\ a \cdot x + b = y(x) \end{cases}$$

Subtraindo-se a 2ª da 1ª equação e a 3ª da 2ª tem-se

$$\begin{cases} a \cdot (x_1 - x_0) = y_1 - y_0 \\ a \cdot (x - x_1) = y(x) - y_1 \end{cases}$$

Ou seja

$$\begin{cases} a = \dfrac{y_1 - y_0}{x_1 - x_0} \\ a = \dfrac{y(x) - y_1}{x - x_1} \end{cases}$$

Então

$$\frac{y_1 - y_0}{x_1 - x_0} = \frac{y(x) - y_1}{x - x_1}$$

Segue que

$$y(x) - y_1 = (x - x_1)\frac{y_1 - y_0}{x_1 - x_0}$$

Logo

$$y(x) = (x - x_1)\frac{y_1 - y_0}{x_1 - x_0} + y_1$$

De onde

$$y(x) = (x - x_1)\frac{y_1 - y_0}{x_1 - x_0} + y_1$$

Assim

$$y(x) = (x - x_1)\frac{y_1 - y_0}{x_1 - x_0} + y_1 \cdot \frac{(x_1 - x_0)}{(x_1 - x_0)}$$

Então

$$y(x) = \frac{y_1 \cdot (x - x_1) - y_0 \cdot (x - x_1)}{x_1 - x_0} + \frac{y_1 \cdot (x_1 - x_0)}{x_1 - x_0}$$

Segue

$$y(x) = \frac{x_1 - x}{x_1 - x_0} \cdot y_0 + \frac{x - x_0}{x_1 - x_0} \cdot y_1$$

Finalmente

$$y(x) = \frac{x - x_1}{x_0 - x_1} \cdot y_0 + \frac{x - x_0}{x_1 - x_0} \cdot y_1 \tag{4.2}$$

Exemplo 4.1

Encontrar uma aproximação para $sen(0.07)$ por interpolação linear da tabela:

x	sen(x)
0.05	0.0499792
0.09	0.0898785

Substituindo o par de pontos $(0.05,\ 0.0499792)$ e $(0.09,\ 0.0898785)$ e colocando o valor $x = 0.07$ na equação 4.2:

$$y(0.07) = \frac{0.07 - 0.09}{0.05 - 0.09} \cdot 0.0499792 + \frac{0.07 - 0.05}{0.09 - 0.05} \cdot 0.0898785 = 0.0699289$$

Portanto $sin(0.07) = 0.0699289$, aproximado por interpolação linear $y(0.07)$. É bom lembrar que $sex(x) \cong x$ para x próximo de 0. O valor $sen(0.07) = 0.0699428$ é obtido com a função da biblioteca SCILAB. Aqui estão os cálculos exibidos no console do SCILAB:

```
--> (0.07 - 0.09)/(0.05 - 0.09) * 0.0499792    + (0.07 - 0.05)/(0.09 - 0.05) * 0.0898785
 ans  =

    0.0699289

-->sin(0.07)
 ans  =

    0.0699428
```

-->

Exemplo 4.2

Encontrar uma aproximação para $y(2)$ por interpolação linear da tabela:

x	y
1	3
3	7
5	18

Substituindo o par de pontos $(1,\ 3)$ e $(3,\ 7)$:

$$y(1) = a \cdot 1 + b = 3$$
$$y(3) = a \cdot 3 + b = 7$$

ou seja

$$\begin{cases} 1 \cdot a + b = 3 \\ 3 \cdot a + b = 7 \end{cases} \quad \therefore \quad \begin{array}{l} a = 2 \\ b = 1 \end{array}$$

$$y(x) = 2 \cdot x + 1,\ x \in [1, 3]$$
$$y(2) = 2 \cdot 2 + 1 = 5$$

colocando o valor $x = 2$ na equação.

4.4 Interpolação Quadrática

Denomina-se interpolação quadrática o método de interpolação que se utiliza de uma função quadrática $y(x)$ (um polinômio de segundo grau) para representar, por aproximação, uma função f(x) (ou suposta f(x) que representa um fenômeno real empiricamente observado) contido no domínio de f(x). A interpolação quadrática é uma parábola que se ajusta a três pontos. A interpolação quadrática é mostrada na Figura 2 onde $y_0,\ y_1$ e y_2 são os valores conhecidos de $y(x) = f(x)$ em $x = x_0,\ x = x_1$ e $x = x_2$, respectivamente.

O modelo quadrático é dado por

$$y(x) = a \cdot x^2 + b \cdot x + c \tag{4.3}$$

Substituindo os pontos $(x_0, y_0),\ (x_1, y_1)$ e (x_2, y_2) na equação 4.3 tem-se

$$\begin{cases} a \cdot x_0^2 + b \cdot x_0 + c = y_0 \\ a \cdot x_1^2 + b \cdot x_1 + c = y_1 \\ a \cdot x_2^2 + b \cdot x_2 + c = y_2 \end{cases} \tag{4.4}$$

A solução do sistema linear 4.4 permitem encontrar os parâmetros $a,\ b,\ c$ do modelo dado pela equação 4.3.

Exemplo 4.3

Encontrar uma aproximação para $y(2)$ por interpolação quadrática da tabela:

x	y
1	3
3	7
5	18

Substituindo o par de pontos $(1,\ 3),\ (3,\ 7)$ e $(5,\ 18)$ e colocando o valor $x = 0.07$ na equação :

$$y(1) = a \cdot 1^2 + b \cdot 1 + c = 3$$
$$y(3) = a \cdot 3^2 + b \cdot 3 + c = 7$$
$$y(5) = a \cdot 5^2 + b \cdot 5 + c = 18$$

ou seja

$$\begin{cases} 1 \cdot a + 1 \cdot b + c = 3 \\ 9 \cdot a + 3 \cdot b + c = 7 \\ 25 \cdot a + 5 \cdot b + c = 18 \end{cases} \quad \therefore \quad \begin{array}{l} a = 0.875 \\ b = -1.5 \\ c = 3.625 \end{array}$$

$$y(x) = 0.875 \cdot x?2 - 1.5 \cdot x + 3.625$$
$$y(2) = 0.875 \cdot 2?2 - 1.5 \cdot x + 3.625 = 4.125$$

colocando o valor $x = 2$ na equação.

Exemplo 4.4

Encontrar uma aproximação para $sen(0.07)$ por interpolação linear da tabela:

x	sen(x)
0.05	0.0499792
0.09	0.0898785
0.12	0.1197122

Substituindo os pontos $(0.05,\ 0.0499792)$, $(0.09,\ 0.0898785)$ e $(1.12,\ 0.1197122)$ e colocando o valor $x = 0.07$ na equação 4.4 com um pequeno rearranjo obtém-se:

$$\begin{cases} 0.05^2 \cdot a + 0.05 \cdot b + c = 0.0499792 \\ 0.09^2 \cdot a + 0.09 \cdot b + c = 0.0898785 \\ 1.12^2 \cdot a + 1.12 \cdot b + c = 0.1197122 \end{cases} \tag{4.5}$$

Solucionando 4.5 tem-se: $a = -0.1970648$, $b = 1.0250716$ e $c = -0.0007817$

Colocando o valor $x = 0.07$ na equação 4.3:

$$y(0.07) = a \cdot 0.07^2 + b \cdot 0.07 + c = 0.0700077$$

Portanto $sin(0.07) = 0.0699461$, aproximado por interpolação quadrática $y(0.07)$. É bom lembrar que $sex(x) \cong x$ para x próximo de 0. O $sen(0.07) = 0.0699428$ é obtido com a função da biblioteca SCILAB.

Aqui estão os cálculos exibidos no console do SCILAB:

```
-->A=[ 0.05^2 0.05 1
-->0.09^2 0.09 1
-->0.12^2 0.12 1];

-->b=[0.0499792;  0.0898785; 0.1197122];

-->x=A\b
 x  =

  - 0.0432262
    1.0035342
  - 0.0000894

-->a=x(1);b=x(2);c=x(3);

-->y=a*0.07^2+b*0.07+c
 y  =

    0.0699461

-->sin(0.07)
 ans  =

    0.0699428

-->
```

A resolução do sistema no exemplo acima é um processo bastante simples e exato na obtenção de $y(x)$. Não se pode esperar que isto ocorra sempre para qualquer problema de interpolação. A matriz A dos coeficientes do sistema é uma matriz de Vandermonde, podendo ser mal-condicionada.

4.5 Interpolação Polinomial

Dado os pontos $(x_0,\ y_0)$, $(x_1,\ y_1)$, ..., $(x_n,\ y_n)$, correspondente a $(n+1)$ pontos diferentes. Interpolar significa determinar o valor de y que é a imagem de um valor x, diferente de $x_i,\ i = 1 \cdots n$, tal que $x_0 < x < x_n$. Pelos pontos $P_0,\ P_1,\ \cdots P_n$ é determinado um único polinômio $P_n(x)$ de grau n (Figura 3). Por dois pontos tem-se uma reta; por três pontos não alinhados, tem-se uma parábola; etc. Assim, por $(n+1)$ pontos passa um polinômio de grau n por eles que permite calcular para cada valor de x um valor y que é o valor da interpolação.O modelo polinomial é dado por

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

Existem várias formas de se obter um polinômio interpolador. Uma das formas é a solução do sistema linear de grau $n+1$ para obtenção de um polinômio interpolador $p_n(x)$, de grau menor ou igual a n. Será estudado as formas de Lagrange, Newton, Gregory-Newton. Teoricamente todas a formas de obtenção de um polinômio de grau n conduzem ao mesmo polinômio. A escolha depende de tempo computacional, estabilidade, etc.

Teorema 4.1

Se $x_0, x_1, \cdots, x_n$ são números reais distintos, então para valores arbitrários $y_0, y_1, \cdots, y_n$ respectivamente formando pares ordenados $(x_i, y_i) com (0 \leq i \leq n)$, existe um único polinômio de grau no máximo n tal que $p_n(x_i) = y_i \ \ (0 \leq i \leq n)$.

4.5.1 Matrix de Vandermonde

Seja o modelo polinomial

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

com $p(x_i) = y_i, \quad \forall i \in \{0, 1, \cdots, n\}$. Então

$$\begin{pmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \cdots & x_1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \cdots & x_n & 1 \end{pmatrix} \begin{pmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

A matriz da esquerda é chamada de matriz de Vandermonde. O número de condição da matriz de Vandermonde pode ser grande, acarretando grandes erros nos cálculos do coeficientes a_i se o sistema de equações é solucionado pelo método de eliminação de Gauss.

4.6 Polinômio Interpolador de Lagrange

O polinômio de Lagrange (nome é devido a Joseph-Louis de Lagrange) é um polinômio de interpolação de um conjunto de pontos na forma de Lagrange.

Dado um conjunto de $n+1$ pontos: $(x_0,\ y_0)$, $(x_1,\ y_1)$, ..., $(x_n,\ y_n)$, com todos x_i distintos, o polinômio de interpolação de um conjunto de pontos na forma de Lagrange é a combinação linear dos polinômios da base de Lagrange:

$$P_n(x) = \sum_{i=0}^{n} y_i \cdot L_i(x) \tag{4.6}$$

com polinômios da base de Lagrange dados por:

$$L_i(x) = \prod_{\substack{j \neq i \\ j=0}}^{n} \frac{(x - x_j)}{(x_i - x_j)} = \frac{(x - x_0)}{(x_i - x_0)} \cdots \frac{(x - x_{i-1})}{(x_i - x_{i-1})} \cdot \frac{(x - x_{i+1})}{(x_i - x_{i+1})} \cdots \frac{(x - x_n)}{(x_i - x_n)} \tag{4.7}$$

O polinômio interpolador de Lagrange é dado por:

$$P_n(x) = \sum_{i=0}^{n} y_i \cdot \prod_{\substack{j \neq i \\ j=0}}^{n} \frac{(x - x_j)}{(x_i - x_j)} \tag{4.8}$$

Exemplos de $P_n(x)$:

$$P_1(x) = y_0 \cdot \frac{(x - x_1)}{(x_0 - x_1)} + y_1 \cdot \frac{(x - x_0)}{(x_1 - x_0)}$$

$$P_2(x) = y_0 \cdot \frac{(x - x_1) \cdot (x - x_2)}{(x_0 - x_1) \cdot (x_0 - x_2)} + y_1 \cdot \frac{(x - x_0) \cdot (x - x_2)}{(x_1 - x_0) \cdot (x_1 - x_2)} + y_2 \cdot \frac{(x - x_0) \cdot (x - x_1)}{(x_2 - x_0) \cdot (x_2 - x_1)}$$

$$P_3(x) = y_0 \cdot \frac{(x - x_1) \cdot (x - x_2) \cdot (x - x_3)}{(x_0 - x_1) \cdot (x_0 - x_2) \cdot (x_0 - x_3)} + y_1 \cdot \frac{(x - x_0) \cdot (x - x_2) \cdot (x - x_3)}{(x_1 - x_0) \cdot (x_1 - x_2) \cdot (x_1 - x_3)} + y_2 \cdot \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_3)}{(x_2 - x_0) \cdot (x_2 - x_1) \cdot (x_2 - x_3)} + y_3 \cdot \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_2)}{(x_3 - x_0) \cdot (x_3 - x_1) \cdot (x_3 - x_2)}$$

Temos abaixo a complexidade da interpolação de Lagrange para um $p_n(x)$ com relação ao número de operações:

Operações	Complexidade
Adições	$2n^2 + 3n + 1$
Multiplicações	$2n^2 + 3n + 1$
Divisões	$n + 1$
Total	$4n^2 + 7n + 3$

Detalhe do cálculo de complexidade (calcula-se L_i e depois executa-se o somatório):

Adições:

- são $2n$ adições (diferenças) para cada dos $L_i (= n + 1)$ dando um total de $2n \cdot (n + 1)$;
- segue-se mais $(n + 1)$ adições;
- Total de adições: $2n \cdot (n + 1) + n + 1 = 2n^2 + 3n + 1$

Multiplicações:

- são $2n$ multiplicações para cada dos $L_i (= n + 1)$ dando um total de $2n \cdot (n + 1)$;
- segue-se mais $(n + 1)$ multiplicações;
- Total de multiplicações: $2n \cdot (n + 1) + n + 1 = 2n^2 + 3n + 1$

Divisões:

- é 1 divisão para cada um dos $L_i (= n + 1)$;
- Total de divisões: $n + 1$

Exemplo 4.5

A tabela abaixo mostra valores de x (variável independente) e y (variável dependente):

x	y
0.25	4.3
0.55	6.9
0.85	11.1
1.15	14.7
1.45	15.6
1.75	27.4

Calcule:

a) y(0.40) utilizando o polinômio de Lagrange de primeiro grau;

b) y(1.25) utilizando o polinômio de Lagrange de segundo grau;

c) y(1.25) utilizando o polinômio de Lagrange de terceiro grau.

Solução

a) $P_1(x) = y_0 \cdot \frac{(x - x_1)}{(x_0 - x_1)} + y_1 \cdot \frac{(x - x_0)}{(x_1 - x_0)}$

Escolhe-se 02 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 0.40$ *entre* $[x_0, x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.25*	*4.3*
1	*0.55*	*6.9*

$$P_1(0.40) = 4.3 \cdot \frac{(0.40 - 0.55)}{(0.25 - 0.55)} + 6.9 \cdot \frac{(0.40 - 0.25)}{(0.55 - 0.25)} = 5.6$$

b) $P_2(x) = y_0 \cdot \frac{(x - x_1) \cdot (x - x_2)}{(x_0 - x_1) \cdot (x_0 - x_2)} + y_1 \cdot \frac{(x - x_0) \cdot (x - x_2)}{(x_1 - x_0) \cdot (x_1 - x_2)} +$
$y_2 \cdot \frac{(x - x_0) \cdot (x - x_1)}{(x_2 - x_0) \cdot (x_2 - x_1)}$

Escolhe-se 03 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 1.25$ *entre* $[x_0,\ x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*1.15*	*14.7*
1	*1.45*	*15.6*
2	*1.75*	*27.4*

$$P_2(1.25) = 14.7 \cdot \frac{(1.25 - 1.45) \cdot (1.25 - 1.75)}{(1.15 - 1.45) \cdot (1.15 - 1.75)} + 15.6 \cdot \frac{(1.25 - 1.15) * (1.25 - 1.75)}{(1.45 - 1.15) * (1.45 - 1.75)} +$$
$$. \ 27.4 \cdot \frac{(1.25 - 1.15) \cdot (1.25 - 1.45)}{(1.75 - 1.15) \cdot (1.75 - 1.45)} = 13.8$$

c) $P_3(x) = y_0 \cdot \frac{(x - x_1) \cdot (x - x_2) \cdot (x - x_3)}{(x_0 - x_1) \cdot (x_0 - x_2) \cdot (x_0 - x_3)} +$
$y_1 \cdot \frac{(x - x_0) \cdot (x - x_2) \cdot (x - x_3)}{(x_1 - x_0) \cdot (x_1 - x_2) \cdot (x_1 - x_3)} +$
$y_2 \cdot \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_3)}{(x_2 - x_0) \cdot (x_2 - x_1) \cdot (x_2 - x_3)} +$
$y_3 \cdot \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_2)}{(x_3 - x_0) \cdot (x_3 - x_1) \cdot (x_3 - x_2)}$

Escolhe-se 04 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 1.25$ *entre* $[x_0,\ x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.85*	*11.1*
1	*1.15*	*14.7*
2	*1.45*	*15.6*
3	*1.75*	*27.4*

$$P_3(x) = 11.1 \cdot \frac{(1.25 - 1.15) \cdot (1.25 - 1.45) \cdot (1.25 - 1.75)}{(0.85 - 1.15) \cdot (0.85 - 1.45) \cdot (0.85 - 1.75)} +$$
$$14.7 \cdot \frac{(1.25 - 0.85) \cdot (1.25 - 1.45) \cdot (1.25 - 1.75)}{(1.15 - 0.85) \cdot (1.15 - 1.45) \cdot (1.15 - 1.75)} +$$
$$15.6 \cdot \frac{(1.25 - 0.85) \cdot (1.25 - 1.15) \cdot (1.25 - 1.75)}{(1.45 - 0.85) \cdot (1.45 - 1.15) \cdot (1.45 - 1.75)} +$$
$$27.4 \cdot \frac{(1.25 - 0.85) \cdot (1.25 - 1.15) \cdot (1.25 - 1.45)}{1.75 - 0.85) \cdot (1.75 - 1.15) \cdot (1.75 - 1.45)} = 14.6$$

4.6.1 Polinômio de Lagrange - Fórmula Ninja

Será apresentado agora um esquema prático para calcular o valor do polinômio de interpolação num ponto (não tabelado) de um forma mais econômica.

Consideremos a fórmula de Lagrange, (4.6), e a fórmula dos $L_i(x)$, (4.7). Definindo:

$Prod_x = (x - x_0)(x - x_1) \cdots (x - x_n)$, ,
$Prodi_i = (x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdots (x_i - x_n),\ i = 0 \cdots n,$
$Dif_i = (x - x_i),\ i = 0 \cdots n.$
pode-se escrever:

$$L_i(x) = \frac{Prod_x}{Prod_i \cdot Dif_i} \tag{4.9}$$

onde: $Prod_i$ é a derivada de $Prod_x$ avaliada em $x = x_i$.

Assim, a fórmula de Lagrange pode ser escrita como:

$$P_n(x) = Prod_x \sum_{i=0}^{n} \frac{y_i}{Prod_i \cdot Dif_i}$$

se reduz a

$$P_n(x) = Prod_x \left(\frac{y_0}{Prod_0 \cdot Dif_0} + \frac{y_1}{Prod_1 \cdot Dif_1} + \cdots + \frac{y_n}{Prod_n \cdot Dif_n} \right). \tag{4.10}$$

Temos abaixo a complexidade da interpolação de Lagrange - Fórmula Ninja, para um $p_n(x)$ com relação ao número de operações:

Operações	Complexidade
Adições	$n^2 + 2n + 1$
Multiplicações	$n^2 + 2n + 1$
Divisões	$n + 1$
Total	$2n^2 + 5n + 3$

Detalhe do cálculo de complexidade (calcula-se Π e depois executa-se o somatório):

Adições:

- são $(n + 1)$ adições (diferenças = Dif_i) para $Prod_x$;
- são n adições (diferenças) para cada $Prod_i$, o que dá total $(n + 1) \cdot n$ adições (diferenças) nos cálculos para $Prod_i$;
- Total de adições: $(n + 1) + (n + 1) \cdot n = n^2 + 2n + 1$

Multiplicações:

- são n multiplicações para $Prod_x$;
- segue-se mais $(n - 1)$ multiplicações para cada $Prod_i$, o que dá total $(n - 1) \cdot (n + 1) = n^2 - 1$ multiplicações nos cálculos para os $Prod_i$;
- têm-se $(n + 1)$ multiplicações de $Dif_i \cdot Prod_i$;
- uma última multiplicação $Prod_x \cdot \sum$
- Total de multiplicações: $n + (n^2 - 1) + (n + 1) + 1 = n^2 + 2n + 1$

Divisões:

- é 1 divisão para cada um dos $\frac{y_i}{Prod_i \cdot Dif_i}$;
- Total de divisões: $n + 1$

Exemplo 4.6

A tabela abaixo mostra valores de x (variável independente) e y (variável dependente):

x	y
0.25	4.3
0.55	6.9
0.85	11.1
1.15	14.7
1.45	15.6
1.75	27.4

Calcule:

a) y(0.75) utilizando o polinômio de Lagrange - Fórmula Ninja, de segundo grau;

b) y(0.75) utilizando o polinômio de Lagrange - Fórmula Ninja, de terceiro grau.

Solução

a) $P_2(x) = Prod_x \cdot \left(\frac{y_0}{Prod_0 \cdot Dif_0} + \frac{y_1}{Prod_1 \cdot Dif_1} + \frac{y_2}{Prod_2 \cdot Dif_2} \right)$

Escolhe-se 03 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 0.75$ *entre* $[x_0,\ x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.55*	*6.9*
1	*0.85*	*11.1*
2	*1.15*	*14.7*

Tabela de cálculos:

	$x_0 = 0.55$	$x_1 = 0.85$	$x_2 = 1.15$	Π
$x = 0.75$	*0.20*	-0.10	-0.40	*0.008*
$x_0 = 0.55$	–	-0.30	-0.60	*0.18*
$x_1 = 0.85$	*0.30*	–	*-0.30*	*-0.09*
$x_2 = 1.15$	*0.60*	*0.30*	–	*0.18*

Cálculo da interpolação:

$$P_2(0.75) = 0.008 \cdot \left(\frac{6.9}{(0.18) \cdot (0.20)} + \frac{11.1}{(-0.09) \cdot (-0.10)} + \frac{14.7}{(0.18) \cdot (-0.40)} \right) = 9.8$$

b) $P_3(x) = Prod_x \cdot \left(\frac{y_0}{Prod_0 \cdot Dif_0} + \frac{y_1}{Prod_1 \cdot Dif_1} + \frac{y_2}{Prod_2 \cdot Dif_2} + \frac{y_3}{Prod_3 \cdot Dif_3} \right)$

Escolhe-se 04 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 0.75$ *entre* $[x_0,\ x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.55*	*6.9*
1	*0.85*	*11.1*
2	*1.15*	*14.7*
3	*1.45*	*15.6*

Tabela de cálculos:

	$x_0 = 0.55$	$x_1 = 0.85$	$x_2 = 1.15$	$x_3 = 1.45$	Π
$x = 0.75$	*0.20*	*-0.10*	*-0.40*	*-0.70*	*-0.0056*
$x_0 = 0.55$	–	*-0.30*	*-0.60*	*-0.90*	*-0.162*
$x_1 = 0.85$	*0.30*	–	*-0.30*	*-0.60*	*0.054*
$x_2 = 1.15$	*0.60*	*0.30*	–	*-0.30*	*-0.054*
$x_3 = 1.45$	*0.90*	*0.60*	*0.30*	–	*0.162*

Cálculo da interpolação:

$$P_3(0.75) = -0.0056 \cdot \left(\frac{6.9}{(-0.162)\cdot(0.20)} + \frac{11.1}{(0.054)\cdot(-0.10)} + \frac{14.7}{(-0.054)\cdot(-0.40)} + \frac{15.6}{(0.162)\cdot(-0.70)} \right) = 9.7$$

4.7 Polinômio de Newton

O polinômio de Newton (nome é devido ao físico Isaac Newton), é um polinômio interpolador para um dado conjunto de pontos. Os coeficientes do polinômio são calculados através de diferenças divididas.

A formula ascendente de Newton é obtida de um polinômio que tem a forma abaixo

$$\begin{aligned} P_n(x) = d_0 + (x - x_0)d_1 + (x - x_0)(x - x_1)d_2 + \cdots + \\ + (x - x_0)(x - x_1)\cdots(x - x_{n-1})d_n \end{aligned} \tag{4.11}$$

Este polinômio aproxima uma função desconhecida $y = f(x)$, no intervalo $(x_0,\ x_n)$, com gráfico passando pelos $(n+1)$ pares de coordenadas. Os coeficientes $d_0, d_1, d_2, \cdots, d_n$ são determinados ao se substituir os valores dos pontos na Eq. (4.11), isto é, estabelecendo-se que para $x = x_0, x_1, \cdots, x_n$, $P_n(x)$ é igual a $x = y_0, y_1, \cdots, y_n$, respectivamente. Obtém-se um sistema de equações algébricas

$$\begin{aligned} P_n(x_0) &= y_0 = d_0 \\ P_n(x_1) &= y_1 = d_0 + (x_1 - x_0)d_1 \\ P_n(x_2) &= y_2 = d_0 + (x_2 - x_0)d_1 + (x_2 - x_0)(x_2 - x_1)d_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P_n(x_n) &= y_n = d_0 + (x_n - x_0)d_1 + (x_n - x_0)(x_n - x_1)d_2 + \cdots + \\ &\quad + (x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})d_n \end{aligned} \tag{4.12}$$

As equações (4.12)acima formam um sistema triangular, que pode ser resolvido por substituições sucessivas.

A fórmula do polinômio de Newton dada através de diferenças divididas é:

$$\begin{aligned} P_n(x) = y_0 + (x - x_0)\triangle y_0 + (x - x_0)(x - x_1)\triangle^2 y_0 + \cdots + \\ + (x - x_0)(x - x_1)\cdots(x - x_{n-1})\triangle^n y_0 \end{aligned} \tag{4.13}$$

ou

$$P_n(x) = y_0 + \sum_{i=1}^{n} \left(\prod_{j=0}^{i-1} (x - x_j) \right) \triangle^i y_0 \tag{4.14}$$

Onde $\triangle^i y_0$ representa a diferença dividida de $i-$ésima ordem 0.

4.7.1 Operador de diferença dividida

Sejam os pontos $(x_i,\ y_i)$, $i = 0, 1, 2, \cdots, n$. O operador de diferença dividida $\triangle$ é definido como sendo

- ordem 0: $\triangle^0 y_i = y_i$
- ordem 1: $\triangle^1 y_i = \dfrac{\triangle^0 y_{i+1} - \triangle^0 y_i}{x_{i+1} - x_i} = \dfrac{y_{i+1} - y_i}{x_{i+1} - x_i}$
- ordem 2: $\triangle^2 y_i = \dfrac{\triangle^1 y_{i+1} - \triangle^1 y_i}{x_{i+2} - x_i} = \dfrac{\triangle y_{i+1} - \triangle y_i}{x_{i+2} - x_i}$
- ordem n: $\triangle^n y_i = \dfrac{\triangle^{n-1} y_{i+1} - \triangle^{n-1} y_i}{x_{i+n} - x_i}$

4.7.2 Complexidade da interpolação de Newton

O polinômio de Newton (4.14) pode ser avaliado usando o método de Horner usando a forma seguinte

$$P_n(x) = (\cdots(\triangle^n y_0(x - x_{n-1}) + \triangle^{n-1} y_0)(x - x_{n-2}) + \cdots + \triangle y_0) \tag{4.15}$$

Segue abaixo a complexidade da interpolação de Newton para um polinômio de grau n com relação ao número de operações:

Operações	Complexidade
Adições	$n^2 + 3n$
Multiplicações	n
Divisões	$\frac{n^2+n}{2}$
Total	$\frac{3n^2}{2} + \frac{9n}{2}$

Detalhe do cálculo de complexidade (calcula-se a tabela de diferenças divididas e depois executa-se o somatório):

Adições:

- são $2\frac{n(n+1)}{2} = n^2 + n$ adições (diferenças) na tabela de diferenças divididas, mais n adições e n adições (diferenças) no método de Horner;
- Total de adições: $n^2 + n + n + n = n^2 + 3n$

Divisões:

- $\frac{n(n+1)}{2}$ na tabela de diferenças divididas;
- Total de divisões: $\frac{n^2+n}{2}$

Multiplicações:

- são n multiplicações no método de Horner;
- Total de multiplicações: n

Exemplo 4.7

A tabela abaixo mostra valores de x (variável independente) e y (variável dependente):

x	y
0.25	4.3
0.55	6.9
0.85	11.1
1.15	14.7
1.45	15.6
1.75	27.4

Calcule:

a) y(0.75) utilizando o polinômio de Newton de segundo grau;

b) y(0.75) utilizando o polinômio de Newton de terceiro grau.

Solução

a) $P_2(x) = y_0 + (x - x_0)\Delta y_0 + (x - x_0)(x - x_1)\Delta^2 y_0$

Escolhe-se 03 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 0.75$ *entre* $[x_0, x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.55*	*6.9*
1	*0.85*	*11.1*
2	*1.15*	*14.7*

Tabela de cálculos da Diferenças Divididas:

i	x_i	y_i	Δy_0	$\Delta^2 y_0$
0	*0.55*	*6.9*	$\frac{11.1-6.9}{0.85-0.55}=14$	$\frac{12-14}{1.15-0.55}=-3.33333$
1	*0.85*	*11.1*	$\frac{14.7-11.1}{1.15-0.85}=12$	
2	*1.15*	*14.7*		

Cálculos de diferenças/produtos: (Opcional)

$(x=0.75)$	$x_0=0.55$	$x_1=0.85$
$(x-x_i)$	*0.20*	*-0.10*
$\prod_{j=0}^{i}(x-x_j)$	*0.20*	*-0.02*

Cálculo da interpolação:

$P_2(0.75)=6.9+(0.75-0.55)\cdot 14+(0.75-0.55)(0.75-0.85)\cdot(-3.33)$

$P_2(0.75)=6.9+(0.20)\cdot 14+(-0.02)\cdot(-3.33)=9.8$

b) $P_3(x)=y_0+(x-x_0)\Delta y_0+(x-x_0)(x-x_1)\Delta^2 y_0+(x-x_0)(x-x_1)(x-x_2)\Delta^3 y_0$

Escolhe-se 04 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x=0.75$ *entre* $[x_0,\ x_1]$ *(mais próximo possível):*

i	x_i	y_i
0	*0.55*	*6.9*
1	*0.85*	*11.1*
2	*1.15*	*14.7*
3	*1.45*	*15.6*

Tabela de cálculos da Diferenças Divididas:

i	x_i	y_i	Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$
0	*0.55*	*6.9*	$\frac{11.1-6.9}{0.85-0.55}=14$	$\frac{12-14}{1.15-0.55}=-3.33333$	$\frac{-15-(-3.33333)}{1.45-0.55}=-12.9630$
1	*0.85*	*11.1*	$\frac{14.7-11.1}{1.15-0.85}=12$	$\frac{3-12}{1.45-0.85}=-15$	
2	*1.15*	*14.7*	$\frac{15.6-14.7}{1.45-1.15}=3$		
3	*1.45*	*15.6*			

Cálculos de diferenças/produtos:(Opcional)

$(x=0.75)$	$x_0=0.55$	$x_1=0.85$	$x_2=1.15$
$(x-x_i)$	*0.20*	*-0.10*	*-0.40*
$\prod_{j=0}^{i}(x-x_j)$	*0.20*	*-0.02*	*0.008*

Cálculo da interpolação:

$P_3(0.75)=6.9+(0.75-0.55)\cdot 14+(0.75-0.55)(0.75-0.85)\cdot(-3.33)+(0.75-0.55)(0.75-0.85)(0.75-1.15)\cdot(-12.96)$

$P_3(0.75)=6.9+(0.20)\cdot 14+(-0.02)\cdot(-3.33)+(0.008)\cdot(-12.96)=9.7$

4.8 Polinômio de Gregory-Newton

Quando se conhece $y_0, y_1, y_2, \cdots, y_n$, que correspondentes aos $(n+1)$ valores igualmente espaçados das abscissas, a formula de Newton pode ser simplificada, resultando na formula de Gregory-Newton. Portanto, o polinômio de Gregory-Newton é um caso particular do polinômio de Newton para pontos igualmente espaçados.

Usando a notação de diferenças finita ascendente com o operador Δ, obtém-se a fórmula do polinômio de Gregory-Newton de grau n

$$P_n(z) = y_0 + \frac{z}{1!}\Delta y_0 + \frac{z(z-1)}{2!}\Delta^2 y_0 + \cdots + \frac{z(z-1)\cdots[z-(n-1)]}{n!}\Delta^n y_0 \tag{4.16}$$

ou

$$P_n(x) = y_0 + \sum_{i=1}^{n} \left(\frac{\prod_{j=0}^{i-1}(z-j)}{j!} \right) \Delta^i y_0 \tag{4.17}$$

onde $\Delta^i y_0$ representa a diferença dividida 0 de $i-$ésima ordem, $\Delta^0 y_0 = y_0 = P_n(x_0)$ e utilizou-se a variável auxiliar $z = \dfrac{x - x_0}{h}$, sendo $x_{i+1} - x_i = h,\ \forall i$.

4.8.1 Operador de diferença dividida

Sejam os pontos $(x_i,\ y_i)\ ,\ i = 0, 1, 2, \cdots, n$, sendo $x_{i+1} - x_i = h,\ \forall i$. O operador diferença finita ascendente Δ é definido como sendo

- ordem 0: $\Delta^0 y_i = y_i$
- ordem 1: $\Delta^1 y_i = \Delta^0 y_{i+1} - \Delta^0 y_i = y_{i+1} - y_i$
- ordem 2: $\Delta^2 y_i = \Delta^1 y_{i+1} - \Delta^1 y_i = \Delta y_{i+1} - \Delta y_i$
- ordem n: $\Delta^n y_i = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i$

4.8.2 Complexidade da interpolação de Gregory-Newton

O polinômio de Gregory-Newton (4.16) pode ser avaliado usando o método de Horner usando a forma seguinte

$$\begin{aligned} P_n(x) = (((\cdots(\triangle^n y_0 \frac{(z-n+1)}{n} + \triangle^{n-1} y_0)\frac{(z-n+2)}{n-1} + \cdots + \\ + \triangle^2 y_0)\frac{(z-1)}{2} + \triangle y_0)\frac{(z-0)}{1} + y_0) \end{aligned} \tag{4.18}$$

Tem-se abaixo a complexidade da interpolação de Gregory-Newton para um polinômio de grau n com relação ao número de operações:

Operações	Complexidade
Adições	$\frac{1}{2}n^2 + \frac{7}{2}n + 2$
Multiplicações	n
Divisões	$n+1$
Total	$\frac{n^2}{2} + \frac{11n}{2} + 3$

Exemplo 4.8

A tabela abaixo mostra valores de x (variável independente) e y (variável dependente):

x	y
0.25	4.3
0.55	6.9
0.85	11.1
1.15	14.7
1.45	15.6
1.75	27.4

Calcule:

a) y(1.25) utilizando o polinômio de Gregory-Newton de segundo grau;

b) y(1.25) utilizando o polinômio de Gregory-Newton de terceiro grau.

Solução:

a) $P_2(z) = y_0 + \frac{z}{1!} \cdot \triangle y_0 + \frac{z(z-1)}{2} \triangle^2 y_0$

Escolhe-se 03 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 1.25$ entre $[x_0,\ x_1]$ (mais próximo possível):

i	x_i	y_i
0	1.15	14.7
1	1.45	15.6
2	1.75	27.4

Tabela de cálculos da Diferenças Finitas:

i	y_i	$\triangle y_0$	$\triangle^2 y_0$
0	14.7	$15.6 - 14.7 = 0.90$	$11.8 - 0.9 = 10.90$
1	15.6	$27.4 - 15.6 = 11.8$	–
2	27.4	–	–

Cálculo da interpolação:

$h = 0.30$;

$x = 1.25$;

$x_0 = 1.15$;

$z = \frac{1.25 - 1.15}{0.30} = 0.33333333$;

então

$P_2(0.33) = 14.7 + (0.33) \cdot 0.90 + (0.33)(0.33 - 1) \cdot (10.90) = 13.8$

$P_2(0.33) = 13.8 = y(1.25)$

b) $P_3(z) = y_0 + \frac{z}{1!} \cdot \triangle y_0 + \frac{z(z-1)}{2!}\triangle^2 y_0 + \frac{z(z-1)(z-2)}{3!}\triangle^3 y_0$

Escolhe-se 04 pontos consecutivos e coloca-se, se possível, o valor a ser interpolado $x = 1.25$ entre $[x_0,\ x_1]$ (mais próximo possível):

i	x_i	y_i
0	0.85	11.1
1	1.15	14.7
2	1.45	15.6
3	1.75	27.4

Tabela de cálculos da Diferenças Finitas:

i	y_i	$\triangle y_0$	$\triangle^2 y_0$	$\triangle^3 y_0$
0	11.1	$14.7 - 11.1 = 3.6$	$0.9 - 3.6 = -2.7$	$10.9 - (-2.7) = 13.6$
1	14.7	$15.6 - 14.7 = 0.9$	$11.8 - 0.9 = 10.9$	–
2	15.6	$27.4 - 15.6 = 11.8$	–	–
3	27.4	–	–	–

Cálculo da interpolação:

$h = 0.30;$
$x = 1.25;$
$x_0 = 0.85;$
$z = \frac{1.25 - 0.85}{0.30} = 1.33333333;$
então

$$P_3(1.33) = 11.1 + (1.33) \cdot 3.6 + \frac{(1.33)(1.33-1)}{2} \cdot (-2.7) +$$

$$\frac{(1.33)(1.33-1)(1.33-2)}{2} \cdot (13.6)$$

$$P_3(1.33) = 14.6 = y(1.25)$$

4.8.3 Erro de Interpolação

Se $f(x)$ é $n+1$ vezes continuamente diferenciável em um intervalo fechado $[a,b]$ e $p_n(x)$ é um polinômio de grau no máximo n que interpola $f(x)$ em $n+1$ pontos distintos $\{(x_i, y_i),\ 0 \leq i \leq n\}$ nesse intervalo, então para cada x no intervalo existe ξ nesse intervalo tal que

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

.

O limite de erro acima sugere a escolha dos pontos de interpolação x_i de forma que o produto $\left|\prod (x - x_i)\right|$, é o menor possível. Os nós Chebyshev obtém isso.

Para intervalos igualmente espaçados No caso dos valores de interpolação igualmente espaçados na variável independente, onde $x_i = a + i \cdot h, \forall i = \{0, 1, \ldots, n\}$ e onde $h = \frac{b-a}{n}$, o limite de erro pode ser dado como

$$|f(x) - p_n(x)| \leq \frac{h^{n+1}}{4(n+1)} \max_{\xi \in [a,b]} \left|f^{(n+1)}(\xi)\right|$$

No entanto, foi assumido que $f^{(n+1)}(\xi)$ é dominado por h^{n+1}, ou seja, $f^{(n+1)}(\xi) \cdot h^{n+1} \ll 1$. Acontece casos o erro aumenta à medida que $n \to \infty$ (fenômeno de Runge).

Nota *Isto é utilizado para encontrar limites de erros quando se usa polinômios para aproximação de funções.*

Capítulo Execícios

1. A tabela abaixo mostra valores de $f(x) = x\sqrt{x}$:

x	f(x)
2	2.83
2.5	3.95
3.2	5.72
3.9	7.70
4.1	8.30
5.0	11.18

Calcule:
a) f(3.5) utilizando o polinômio de Lagrange de primeiro grau;
b) f(4.0) utilizando o polinômio de Lagrange de segundo grau;
c) f(3.0) utilizando o polinômio de Lagrange de terceiro grau.

2. A tabela abaixo mostra valores de $f(x) = x\sqrt{x}$:

x	f(x)
2	2.83
2.5	3.95
3.2	5.72
3.9	7.70
4.1	8.30
5.0	11.18

Calcule:

a) f(3.0) utilizando o polinômio de Lagrange com a Fórmula-Ninja de primeiro grau;
b) f(3.5) utilizando o polinômio de Lagrange com a Fórmula-Ninja de segundo grau;
c) f(4.0) utilizando o polinômio de Lagrange com a Fórmula-Ninja de terceiro grau.

3. Dada a tabela abaixo :

x	f(x)
2	2.83
2.5	3.95
3.0	5.72
3.5	7.70
4.0	8.30
4.5	11.18

Calcule:
a) f(2.3) utilizando o polinômio de Newton de primeiro grau;
b) f(2.7) utilizando o polinômio de Newton de segundo grau;
c) f(3.7) utilizando o polinômio de Newton de terceiro grau.

4. Dada a tabela abaixo :

x	f(x)
2	4.83
2.5	5.95
3.0	7.72
3.5	9.70
4.0	12.30
4.5	15.18

Calcule:

a) f(2.3) utilizando o polinômio de Gregory-Newton de primeiro grau;
b) f(2.7) utilizando o polinômio de Gregory-Newton de segundo grau;
c) f(3.7) utilizando o polinômio de Gregory-Newton de terceiro grau.

Capítulo - Integração

A integral definida

$$I = \int_a^b f(x)dx = F(a) - F(b) = \text{número, onde } \frac{dF(x)}{dx} = f(x)$$

.

Os métodos numéricos que serão apresentados para calcular aproximações da integral definida serão fórmulas de integração do tipo:

$$I_Q = \sum_{j=1}^{k} a_i \cdot f(x_i)$$

também designadas fórmulas de quadratura. Os pontos $x_0 = a < x_1 ... < x_n = b$ são os nós de integração e os pesos $a_0, \cdots, a_n$ são coeficientes a determinar e independem da função $f(x)$.

No ramo matemático da análise real, a integral de Riemann define de forma rigorosa uma integral de uma fução definida em um intervalo. Encontra-se a area exata sobre a curva,

$$I = \int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x$$

sob as seguintes condições

1. $f(x)$ é continua no intervalo$[a, b]$;
2. o intervalo $[a, b]$ é dividido em n sub-intervalos de larguras iguais a $\Delta x = \frac{b-a}{n}$;
3. o extremos destes intervalos são os nós de integração $x_0 = a < x_1 ... < x_n = b$;
4. $x_0^*, x_1^*, \cdots, x_n^*$ são qualquer ponto amostal nestes sub-intervalos.

Se o limite acima existe, função $f(x)$ é dita ser Riemann integrável no intervalo fechado [a,b] e a integral definida exite. Observe que os pontos amostrais x_i^* podem ser qualquer ponto amostral no *i-ésimo* sub-intervalo, isto inclui os pesos das formulas de quadraturas que estão associadas a soma de Riemann.

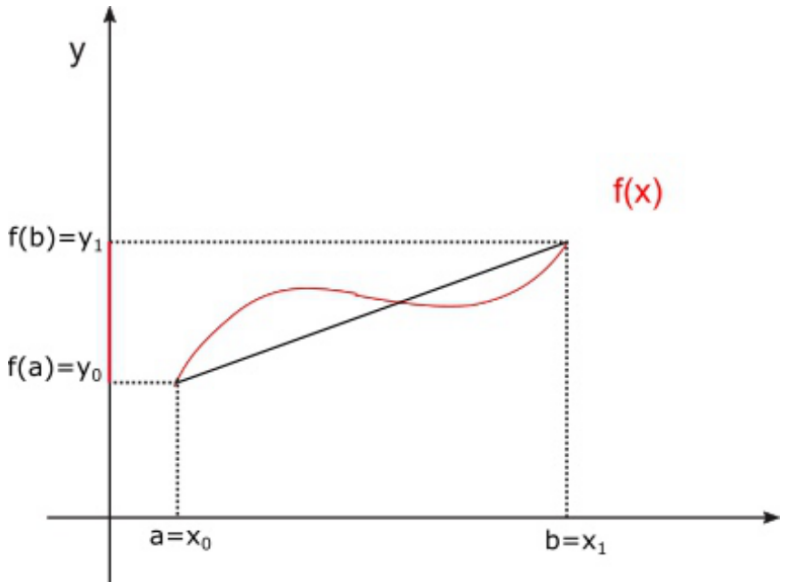


Figura 5.1: Aproximação de uma área por um trapézio.

5.1 Regra dos Trapézios

A Regra dos Trapézios é um método muito simples de aproximar uma integral definida através da área sob uma curva aproximada por uma série de trapézios. A figura 5.1 a apresenta o caso de um intervalo

$$I_T = \frac{(b-a)}{2}\left[f(a)+f(b)\right] = \frac{h}{2}\left(y_0+y_1\right), \quad h = b-a.$$

Claramente comete-se um erro apreciável e para melhorar a exatidão da aproximação pode-se dividir a área e aproximá-la por n trapézios. Teoricamente, usando-se um número infinito de trapézios, teremos a o valor da integral exata, mas a propagação gera problemas e a partir de um dado número de subdivisões o erro de amostragem faz o erro aumentar.

Para usar a regra dos trapézios divide-se o intervalo $[a,b]$ em n partes iguais, como mostra a figura 5.2. O *i-ésimo* trapézios fica entre $[x_{i-1}, x_i]$, $i = 1\cdots n$ com base $h = \dfrac{b-a}{n}$, altura do lado esquerdo $y_{i-1} = f(x_{i-1})$, e a altura do lado direito é $y_i = f(x_{i-})$. Portanto a *i-ésima* área é,

$$(I_T)_i = \frac{h}{2}\left(y_{i-1}+y_i\right)$$

A área total dos n trapézios, denotada como $I_T(n)$, ou simplesmente I_T:

$$\begin{aligned} I_T(n) &= I_T(1) + I_T(2) + \cdots + I_T(n-1) + I_T(n) \\ &= \frac{h}{2}\left(y_0+y_1\right) + \frac{h}{2}\left(y_1+y_2\right) + \cdots + \frac{h}{2}\left(y_{n-1}+y_n\right) \end{aligned}$$

Então,

$$I_T = \frac{h}{2}\left(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n\right)$$

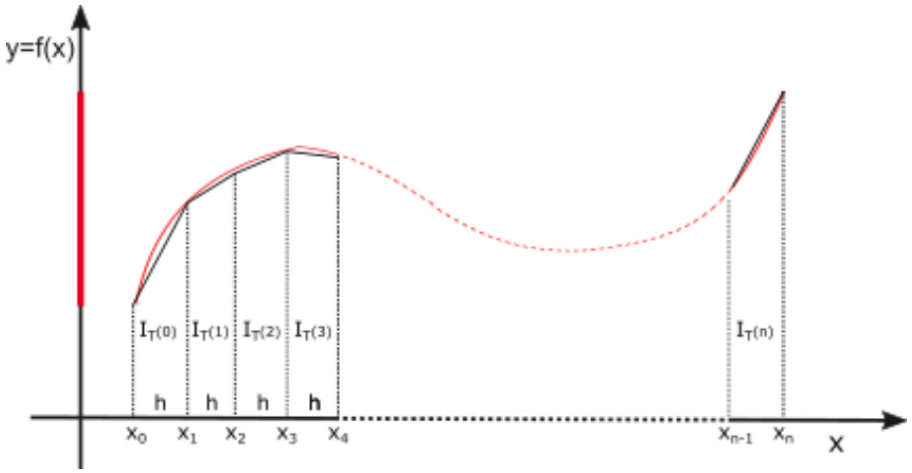


Figura 5.2: Aproximação de uma área por n trapézios.

Exemplo 5.1

Dado que $\int_{-1}^{3} e^x dx = e^3 - e^{-1}$: a) Estimar o valor de usando 2, 4 e 8 sub-intervalos através da Regra dos Trapézios; b) Calcule os erros absoluto e relativo dos valores estimados.

Solução

a) n=2

$$I_T = \frac{h}{2}\left(y_0 + 2y_1 + y_2\right)$$

$$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{2} = 2$$

$$x_i = a + i\ \cdot h,\ \ i = 0\cdots 2,\ f(x) = e^x$$

Tabela 5.1: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	1	2.718
2 = n	3 = b	20.086

$$I_T = \frac{2}{2}\left(0.368 + 2 \cdot 2.718 + 20.086\right) = 25.890$$

n=4

$I_T = \frac{h}{2}\left(y_0 + 2y_1 + 2y_2 + 2y_3 + y_4\right)$

$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{4} = 1$

$x_i = a + i\ \cdot h,\ \ i = 0 \cdots 4,\ f(x) = e^x$

Tabela 5.2: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	0	1
2	1	2.718
3	2	7.389
4=n	3= b	20.086

$$I_T = \frac{1}{2}\left(0.368 + 2 \cdot 1 + 2 \cdot 2.718 + 2 \cdot 7.389 + 20.086\right) = 21.334$$

n=8

$I_T = \frac{h}{2}\left(y_0 + 2y_1 + 2y_2 + 2y_3 + 2y_4 + 2y_5 + 2y_6 + 2y_7 + y_8\right)$

$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{8} = 0.5$

$x_i = a + i\ \cdot h,\ \ i = 0 \cdots 8,\ f(x) = e^x$

Tabela 5.3: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	-0.5	0.606
2	0	1
3	0.5	1.649
4	1	2.718
5	1.5	4.482
6	2	7.389
7	2.5	12.182
8=n	3= b	20.086

$$I_T = \frac{0.5}{2}\left(0.368 + 2 \cdot 0.606 + 2 \cdot 1 + 2 \cdot 1.649 + 2 \cdot 2.718 + 2 \cdot 4.482 + 2 \cdot 7.389 + 2 \cdot 12.182 + 20.086\right) = 20.126$$

b)

$I = \int_{-1}^{3} e^x dx = e^3 - e^{-1} = 19.71765748$

Erro Absoluto: $EA = |I_T - I|$

Erro Relativo(%): $ER(\%) = \frac{EA}{I} \cdot 100\ \%$

Tabela 5.4: Cálculos dos erros para Regra dos Trapézios ($I = 19.718$)

n	I_T	EA	$ER(\%)$
2	25.890	6.172	23.83
4	21.334	1.616	7.57
8	20.126	0.408	2.03

5.2 1^a Regra de Simpson

A regra dos trapézios aproxima a curva a ser integrada por n segmentos de reta como mostram as figuras 5.1 e 5.2. Na 1^a Regra de Simpson a curva é aproximada por uma série de parábolas para aproximar a curva $f(x)$. A figura 5.3 mostra a parábola $y(x) = ax^2 + bx + c$ e a curva $f(x)$. Os pontos $x_0 = a$, $x_2 = b$ e $x_1 = (a + b)/2$ definem a única parábola que passa pelos três pontos que têm coordenadas $(x_0, y_0), (x_1, y_1)$e(x_2, y_2).

A figura 5.3 a apresenta o caso de uma parábola com 2 intervalos iguais $h = \dfrac{b-a}{2}$

$$I_S = \frac{h}{3}\left(y_0 + 4y_1 + y_2\right)$$

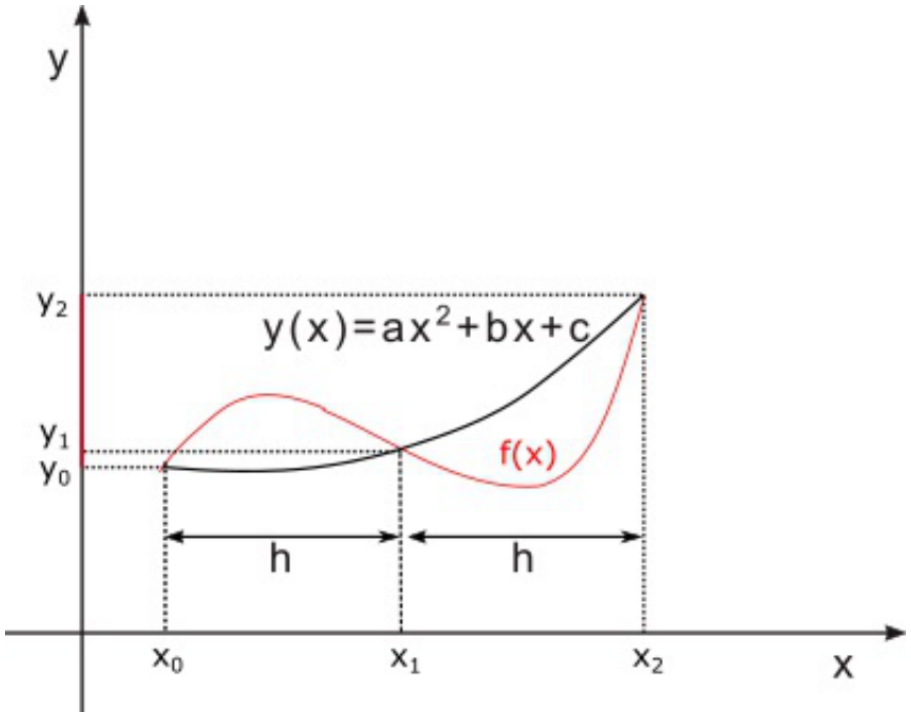


Figura 5.3: Aproximação de uma área por parábola.

Para melhorar a exatidão da aproximação pode-se dividir a área e aproximá-la por n parábolas. Como na regra dos trapézios, teoricamente, usando-se um número infinito de parábolas, teremos a o valor da integral exata, mas a propagação gera problemas e a partir de um dado número de subdivisões o erro de amostragem faz o erro aumentar. Para usar a 1^a regra de Simpson divide-se o intervalo $[a, b]$ em um número par de n partes iguais, como mostra a figura 5.4. A *i-ésimo* parábola passa pelos pontos $(x_{i-2}, y_{i-2}), (x_{i-1}, y_{i-1}),$ e (x_i, y_i), $i = 2, 4, \cdots n$ (um número par), com $h = \dfrac{b-a}{n}$ e $y_i = f(x_i)$. Portanto a *i-ésima* área é,

$$\left(I_S\right)_i = \frac{h}{3}\left(y_{i-2} + 4y_{i-1} + y_i\right).$$

A área total das $n/2$ parábolas (n, número par de intervalos), denotada como $I_S(n)$, ou simplesmente I_S:

$$\begin{aligned} I_S(n) &= I_S(1) + I_S(2) + \cdots + I_S(n/2 - 1) + I_S(n/2) \\ &= \frac{h}{3}\left(y_0 + 4y_1 + y_2\right) + \frac{h}{3}\left(y_2 + 4y_3 + y_4\right) + \cdots + \frac{h}{3}\left(y_{n-2} + 4y_{n-1} + y_n\right) \end{aligned}$$

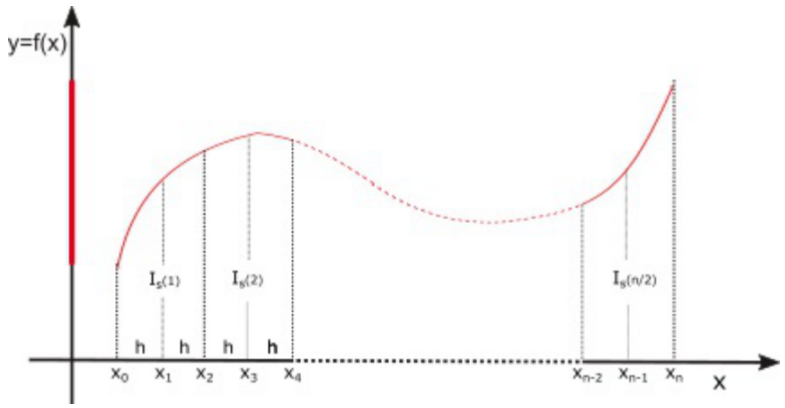


Figura 5.4: Aproximação de uma área por n parábolas.

Então,

$$I_S(n) = \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 4y_{n-1} + y_n\right)$$

Exemplo 5.2

. Dado que $\int_{-1}^{3} e^x dx = e^3 - e^{-1}$: a) Estimar o valor de usando 2, 4 e 8 sub-intervalos através da 1^a Regra de Simpson; b) Calcule os erros absoluto e relativo dos valores estimados.

Solução:

a)

n=2

$I_S = \frac{h}{3}\left(y_0 + 4y_1 + y_2\right)$

$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{2} = 2$

$x_i = a + i\ \cdot h,\ \ i = 0 \cdots 2, f(x) = e^x$

Tabela 5.5: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	1	2.718
2 = n	3 = b	20.086

$$I_S = \frac{2}{3}\left(0.368 + 4 \cdot 2.718 + 20.086\right) = 20.884$$

n=4

$I_S = \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + y_4\right)$

$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{4} = 1$

$x_i = a + i\ \cdot h,\ \ i = 0 \cdots 4, f(x) = e^x$

Tabela 5.6: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	0	1
2	1	2.718
3	2	7.389
4=n	3= b	20.086

$$I_S = \frac{1}{3}\left(0.368 + 4 \cdot 1 + 2 \cdot 2.718 + 4 \cdot 7.389 + 20.086\right) = 19.815$$

n=8

$I_S = \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + 4y_5 + 2y_6 + 4y_7 + y_8\right)$

$a = -1,\ b = 4, \quad h = \frac{b-a}{n} = \frac{3-(-1)}{8} = 0.5$

$x_i = a + i\ \cdot h,\ \ i = 0 \cdots 8, f(x) = e^x$

Tabela 5.7: Cálculos dos valores

i	$x_{[i]}$	$y_i = f(x_{[i]})$
0	-1 = a	0.368
1	-0.5	0.606
2	0	1
3	0.5	1.649
4	1	2.718
5	1.5	4.482
6	2	7.389
7	2.5	12.182
8=n	3= b	20.086

$$I_S = \frac{0.5}{2}\left(0.368 + 4 \cdot 0.606 + 2 \cdot 1 + 4 \cdot 1.649 + 2 \cdot 2.718 + 4 \cdot 4.482 + 2 \cdot 7.389 + 4 \cdot 12.182 + 20.086\right) = 19.724$$

b)

$I = \int_{-1}^{3} e^x dx = e^3 - e^{-}1 = 19.71765748$

Erro Absoluto: $EA = |I_S - I|$

Erro Relativo(%): $ER(\%) = \frac{EA}{I} \cdot 100\ \%$

Tabela 5.8: Cálculos dos erros para 1^a Regra de Simpson ($I = 19.718$)

n	I_S	EA	$ER(\%)$
2	20.884	1.116	5.91
4	19.815	0.097	0.49
8	19.724	0.006	0.030

5.3 Quadratura de Newton-Cotes com Erros

As fórmulas de Newton-Cotes, também chamadas de Quadratura de Newton-Cotes, ou Regras de Newton-Cotes, são um grupo de fórmulas para integração numérica obtidas usando aproximação de $f(x) \approx P_n(x)$ em pontos equidistantes no intervalo $[a, b]$.

As fórmulas de Newton-Cotes podem ser úteis se o valor do integrante, em pontos igualmente espaçados, é fornecido. Se for possível calcular f(x), então outros métodos, como Quadratura Gaussiana e Quadratura de Clenshaw-Curtis são geralmente utilizados.

Assume-se que o valor da função $f(x)$, definida entre $[a, b]$ é calculável em pontos x_i entre $i = 0, \cdots, n$ igualmente espaçados, onde $x_0 < x_1 < \cdots < xn \in [a, b]$. Existem dois tipos de Fórmulas Newton-Cotes, as "fechadas"quando $x_0 = a$ e $x_n = b$, e as "abertas"quando $x_0 > a$ e $x_n < b$.

5.3.1 Fórmulas fechadas

Assumindo que $f(x)$ é tem $n + 1$ derivadas temos a aproximação por interpolação

$$f(x) = P_n(x) + Erro_n(x)$$

onde

$$Erro_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)\frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad x_0 \leq \xi \leq x_n$$

Então

$$\int_{x_0}^{x_n} f(x)dx \approx \int_{x_0}^{x_n} P_n(x)dx$$

e o erro

$$\int_{x_0}^{x_n} Erro_n(x)dx$$

Do polinômio interpolador de Gregory-Newton dado por

$$P_n(z) = y_0 + \frac{z}{1!}\Delta y_0 + \frac{z\cdot(z-1)}{2!}\Delta^2 y_0 + \cdots + \frac{z\cdot(z-1)\cdots(z-[n-1])}{n!}\Delta^n y_0$$

$$z = \frac{x - x_0}{h};$$
$$h = x_i - x_{i-1},\ i = 1\cdots n, \text{(são pontos equidistantes).}$$

e do erro de interpolação na variável z

$$Erro_n(z) = \frac{z(z-1)(z-2)\cdots(z-n)}{(n+1)!}h^{(n+1)}f^{(n+1)}(\xi), \quad x_0 \le \xi \le x_n$$

Então

$$Erro_n(x) = h\int_0^n \frac{z(z-1)(z-2)\cdots(z-n)}{(n+1)!}h^{(n+1)}f^{(n+1)}(\xi)dz$$

para algum ponto $\xi \in [a,\ b]$.

Regra dos Trapézios com Erro

Com um intervalo ($n = 1$)

$$\int_{a=x_0}^{b=x_1} f(x)dx = \frac{h}{2}(y_0 + y_1) - \frac{h^3}{12}f''(\xi), \quad x_0 = a \le \xi \le x_1 = b$$

e com n subintervalos

$$\int_{a=x_0}^{b=x_n} f(x)dx = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n) - \frac{(b-a)}{12}h^2 f''(\xi),$$

$$x_0 = a \le \xi \le x_n = b.$$

1^a Regra de Simpson com Erro

Com uma parábola ($n = 2$)

$$\int_{a=x_0}^{b=x_2} f(x)dx = \frac{h}{3}(y_0 + 4y_1 + y_2) - \frac{h^5}{90}f^{IV}(\xi), \quad x_0 = a \le \xi \le x_2 = b$$

e com n parábolas

$$\int_{a=x_0}^{b=x_n} f(x)dx = \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 4y_{n-1} + y_n) - \frac{(b-a)}{180}h^4 f^{IV}(\xi),$$

$$x_0 = a \le \xi \le x_n = b$$

5.3.2 Limitante superior para o erro

Regra dos Trapézios

Com um intervalo ($n = 1$)

$$|Erro_1| \le \frac{h^3}{12}\text{máx}|f''(x)|,\ \ x_0 = a \le x \le x_1 = b$$

e com n subintervalos

$$|Erro_n| \leq n\frac{h^3}{12}\text{máx}|f''(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

ou

$$|Erro_n| \leq (b-a)\frac{h^2}{12}\text{máx}|f''(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

1^a Regra de Simpson

Com uma parábola ($n = 2$)

$$|Erro_2| \leq \frac{h^5}{90}\text{máx}|f^{IV}(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

e com n (um número par) parábolas

$$|Erro_n| \leq n\frac{h^5}{180}\text{máx}|f'^{IV}(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

ou

$$|Erro_n| \leq (b-a)\frac{h^4}{180}\text{máx}|f^{IV}(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

Exemplo 5.3

Calcule limitante superior das aproximações calculadas pela regra dos trapézios no exemplo5.1.

$$|Erro_n| \leq n\frac{h^3}{12}\text{máx}|f''(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

$$|Erro_2| \leq 2 \cdot \left(\frac{2^3}{12}\right) \cdot 20.086 = 26.78$$

$$|Erro_4| \leq 4 \cdot \left(\frac{1^3}{12}\right) \cdot 20.086 = 6.69$$

$$|Erro_8| \leq 8 \cdot \left(\frac{0.5^3}{12}\right) \cdot 20.086 = 1.67$$

Observa-se que os erros absolutos na tabela **??** estão dentro dos limites calculados acima.

Exemplo 5.4

Calcule limitante superior das aproximações calculadas pela 1^a Regra de Simpson no exemplo 5.2.

$$|Erro_n| \leq n\frac{h^5}{180}\text{máx}|f^{IV}(x)|, \;\; x_0 = a \leq x \leq x_1 = b$$

$$|Erro_2| \leq \left(\frac{2^5}{90}\right) \cdot 20.086 = 7.14$$

$$|Erro_4| \leq 4 \cdot \left(\frac{1^5}{180}\right) \cdot 20.086 = 0.45$$

$$|Erro_8| \leq 8 \cdot \left(\frac{0.5^5}{180}\right) \cdot 20.086 = 0.11$$

Observa-se que os erros absolutos na tabela **??** estão dentro dos limites calculados acima.

Exemplo 5.5

Determinar o número de intervalos que é necessário dividir o intervalo de integração para obter com três casas decimais corretas

$$\int_{-1}^{3} e^x dx$$

pelas: a) Regra dos Trapézios; b) 1^a Regra de Simpson.

Solução

a) Número de intervalos para Regra dos Trapézios

$$|Erro_n| \leq (b-a)\frac{h^2}{12}\textit{máx}|f''(x)|, \;\; -1 \leq x \leq 3$$

$f''(x) = e^x$, *então máx*$|f''(x)| = e^3 = 20.086, \ -1 \leq x \leq 3$

$$4 \times \frac{h^2}{12} \times 20.086 \leq 0.5 \times 10^{-3} \ \ \therefore h^2 \leq 0.0000747 \therefore h \leq 0.00864$$

$$n = \frac{b-a}{h} = \frac{4}{0.00864} \approx 477$$

b) Número de intervalos para 1^a *Regra de Simpson*

$$|Erro_n| \leq (b-a)\frac{h^4}{180} máx|f'^{IV}(x)|, \ -1 \leq x \leq 3$$

$f''(x) = e^x$, *então máx*$|f''(x)| = e^3 = 20.086, \ -1 \leq x \leq 3$

$$4 \times \frac{h^4}{180} \times 20.086 \leq 0.5 \times 10^{-3} \ \ \therefore h^4 \leq 0.001120 \therefore h \leq 0.1829$$

$$n = \frac{b-a}{h} = \frac{4}{0.1829} \approx 22$$

5.3.3 Estimativa da Precisão determinada na prática

A determinação do resultado de uma integral na precisão desejada pode ser obtida usando um procedimento utilizado na prática. Aumenta-se o número de sub-intervalos, fazendo $h \to 0$ ou $n \to \infty$, e comparando-se os resultados,

$$|I_n - I_m| < \epsilon, \quad n > m$$

ou

$$\frac{|I_n - I_m|}{I_n} < \epsilon, \quad n > m$$

O procedimento adotado em cálculos práticos consiste em diminuir h e comparar o resultado assumido ser mais preciso com outro de menor precisão até que um valor pré-fixado é atingido.

Exemplo 5.6

Calcule $\int_{-1}^{3} e^x dx$ usando a 1^a Regra de Simpson com erro estimado menor que 0.01.

Solução

Dos exemplos acima

$I_S(2) = 20.884$
$I_S(4) = 19.815$

Então $|I_S(4) - I_S(2)| = |19.815 - 20.884| = 1.069$ *continuar,*
$I_S(8) = 19.724$

Então $|I_S(8) - I_S(4)| = |19.724 - 19.815| = 0.091$ *continuar,*
$I_S(16) = 19.718$

Então $|I_S(16) - I_S(4)| = |19.718 - 19.724| = 0.006 < \epsilon = 0.01$ *parar.*

Logo $I_S = 19.72 \pm 0.01$.

Capítulo Execícios

1. Calcule

$$I_T = \int_{-2}^{3} \frac{x^4}{1+e^x} dx$$

Usando a Regra dos Trapézios com 8 sub-intervalos.

2. Calcule

$$I_S = \int_{-2}^{3} \frac{x^4}{1+e^x} dx$$

Usando a 1^a Regra de Simpson com 4 sub-intervalos.

3. Faça uma Estimativa da Precisão determinada na prática:

$$Erro = |I_T(\text{questão 1}) - I_S(\text{questão 2})|$$

Capítulo - Ajuste de Curvas

O ajuste de curvas acontece o tempo todo através do desenvolvimento do conhecimento de engenharia, devido ao fato de que muitas vezes temos a capacidade de observar as coisas antes de entendê-las. Assim, o ajuste de curvas quase sempre indica uma solução sub-ótima devido à falta de metodologia ou tempo para soluções mais rigorosos. Mas como projetos de engenharia são complexos, torna-se uma parte regular do pensamento dos engenheiros. A técnica básica de ajuste de curvas de mínimos quadrados será abordada.

Falando mais amplamente sobre a questão do ajuste de curvas é preciso que se entenda que quando as pessoas falam sobre aprendizado de máquina e IA, geralmente estão falando sobre *ajuste de curvas*. As palavras-chave são *regressão*, *aprendizagem supervisionada*, *meta-modelagem*.

6.1 O conceito "ajuste de curva"

O ajuste de curva, ou formalmente, aprendizado supervisionado, lida com dados rotulados $\mathcal{D} = [\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2, \cdots]$ onde $\mathbf{x}$ é a entrada e $\mathbf{y}$ a saída (ou o rótulo).

Por exemplo, as entradas podem ser as declarações de Putin sobre a Ucrânia e as saídas o desempenho do mercado de ações logo em seguida; ou a composição do material de entrada e as saídas sua resistência à fratura; ou as imagens de entrada da câmera e as saídas de reconhecimento de objetos dentro das imagens; ou as entradas de adubos químicos e as saídas da taxa de sobrevivência das pragas; ou ... e assim por diante.

Em aplicações simples, o ajuste de curvas lida com entradas escalares ou vetoriais e saídas escalares. O objetivo é ajustar uma curva (para entrada 1D) ou uma superfície (para entrada 2D) ou uma hipersuperfície (para entrada nD) aos dados para que, quando uma nova entrada for fornecida, possamos encontrar uma previsão precisa da saída correspondente da curva ou superfície ou hipersuperfície. Muitas vezes a pergunta é: usar uma linha reta para ajustar ou algum tipo de curva? Como decidir? Um conceito importante para aprender é que a precisão da previsão de sua curva aumenta quando a curva se ajusta bem aos dados, mas também diminui se sua curva for muito ondulada. Em outras palavras, você deseja encontrar uma curva (por exemplo, uma função polinomial de um certo grau) que se ajuste à maioria dos pontos de dados (um ajuste suficiente), mas não um superajuste. As diversas metodologias utilizadas para este fim incluem critérios de informação de : Akaike (AIC), Akaike Corrigido (AICc), Bayesiano (BIC) e validação cruzada.

Para avaliar as previsões, há duas métricas importantes a serem consideradas: variância e viés.

VARIÂNCIA

A variação é a quantidade pela qual a estimativa da função de destino muda se dados de treinamento diferentes forem usados. A função alvo f estabelece a relação entre as variáveis de entrada (propriedades) e de saída (previsão). Quando um conjunto de dados diferente é usado, a função de destino precisa permanecer estável com pouca variação porque, para qualquer tipo de dado, o modelo deve ser genérico. Para evitar previsões falsas, precisamos garantir que a variância seja baixa. Por essa razão, o modelo deve ser generalizado para aceitar características não vistas de dados de temperatura e produzir melhores previsões.

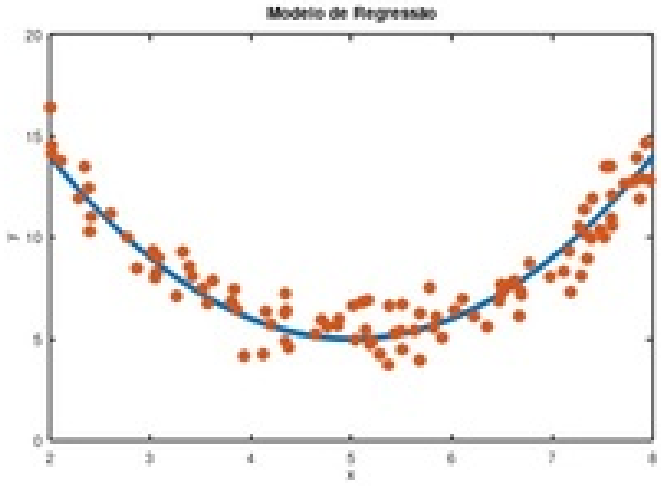
VIÉS

O viés é a tendência do algoritmo de aprender consistentemente a coisa errada por não levar em consideração todas as informações nos dados. Para que o modelo seja preciso, o viés precisa ser baixo. Se houver inconsistências no conjunto de dados, como valores ausentes, menor número de tuplas de dados ou erros nos dados de entrada, o viés será alto e as previsões erradas.

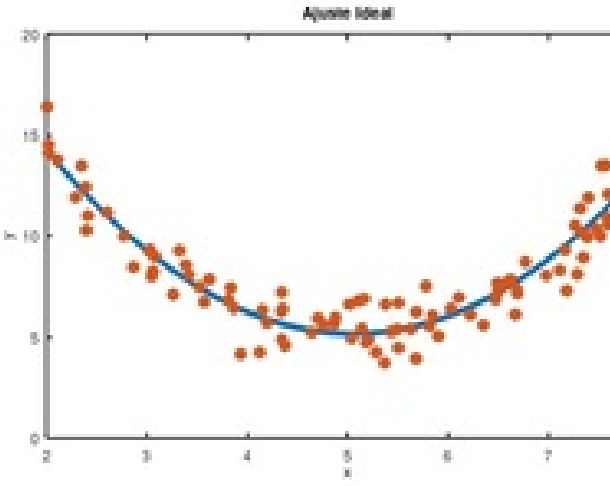
Exatidão e erro são as duas outras métricas importantes. O erro é a diferença entre o valor real e o valor previsto estimado pelo modelo. Exatidão é a fração de previsões certas.

VALIDAÇÃO CRUZADA

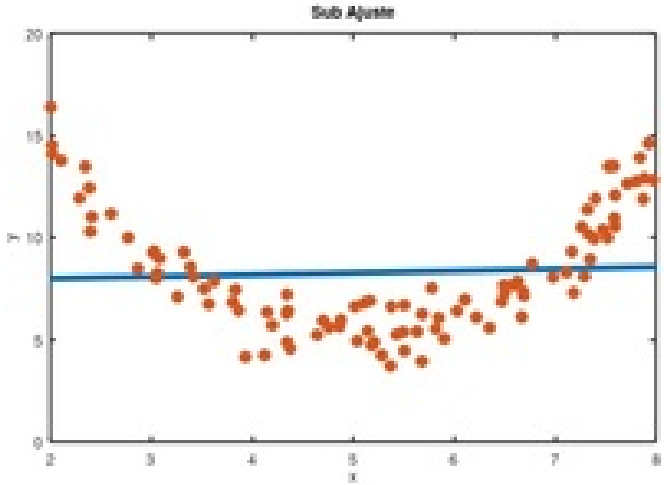
Para que um modelo seja considerado bom, espera-se que tenha baixa variância, baixo viés e baixo erro. Para conseguir isso, precisamos particionar o conjunto de dados em conjuntos de dados de treinamento e teste. O modelo aprenderá padrões do conjunto



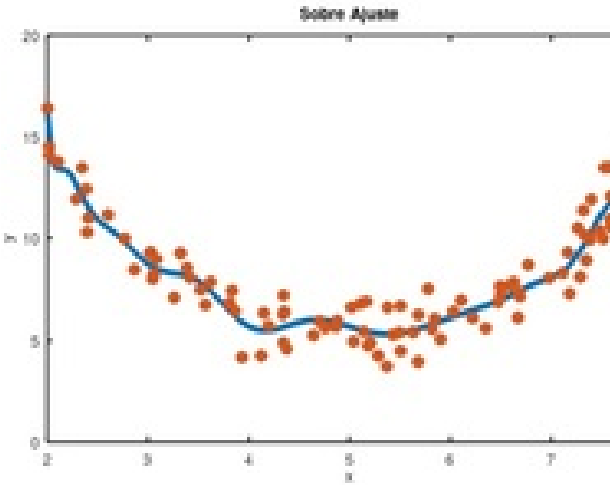
(a) Modelo da Regressão



(b) Ajuste Ideal



(c) Sub Ajuste



(d) Sobre Ajuste

Figura 6.1: Modelo de Regressão e Ajustes Polinomiais.

de dados de treinamento e o desempenho será avaliado no conjunto de dados de teste. Para reduzir o erro enquanto o modelo está aprendendo, utilza-se uma função de erro (função objetiva). Se o modelo memorizar/imitar os dados de treinamento alimentados a ele, em vez de encontrar padrões, ele fornecerá previsões falsas sobre dados não vistos. A curva derivada do modelo treinado passaria por todos os pontos de dados e a precisão no conjunto de dados de teste é baixa. Isso é chamado de *sobreajuste* e irá ser traduzido pela alta variância. Por outro lado, se o modelo tiver um bom desempenho nos dados de teste, mas com baixa precisão nos dados de treinamento, isso levará ao *subajuste* (Ver 6.1 e Apêndice).

Existem vários algoritmos que são usados para construir um modelo de regressão, alguns funcionam bem sob certas restrições e outros não. Aqui não vamos de mergulhar nos algoritmos de ajustes de curva, vamos ver como funciona a técnica básica do ajuste de mínimos quadrados.

6.2 Método dos Mínimos Quadrados

O ajuste de curvas no caso em que se tem uma tabela de pontos $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, com x_i pertencentes ao intervalo $[a, b]$, consiste em dadas $m + 1$ funções $g_0(x), g_1(x), \cdots, g_m(x)$, contínuas em $[a, b]$, determinar $m + 1$ coeficientes $\beta_0, \beta_1, \cdots, \beta_m$ considerando que

$$f(x) = \beta_0 g_0(x) + \beta_1 g_1(x) + \cdots + \beta_m g_m(x)$$

se aproxima de $y(x)$, que fornece os valores $y_0, y_1, ..., y_n$ dos pontos tabelados.

Este é um modelo matemático linear pois os coeficientes β_i a serem determinados aparecem como combinação linear, embora as funções $g_i(x)$ possam ser não lineares, como $g_0(x) = ln(x)$ e g1(x) = cos(x), por exemplo.

O problema é escolher adequadamente estas funções. A observação do diagrama de dispersão pode ser utilizado para ver a forma geral dos pontos, ou então deve-se basear em fundamentos teóricos do experimento que produz a tabela.

Uma técnica para que a função f(x) se ajuste aos pontos y_i é fazer com que o desvio, ou erro i, $d_i^2 = (y_i - f(x_i))^2$ seja mínimo para todo $i = 0, 1, \cdots, n$. Da soma destes desvios elevados ao quadrado tem-se

$$\begin{aligned} D(\beta_0, \beta_1, \cdots, \beta_m) &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} [y_i - f(x_i)]^2 \\ &= \sum_{i=1}^{n} [y_i - \beta_0 g_0(x_i) - \beta_1 g_1(x_i) - \cdots - \beta_m g_m(x_i)]^2 . \end{aligned}$$

e deseja-se determinar os β_i's onde $D(.)$ seja mínimo. Este processo de minimização é chamado de *Método dos Mínimos Quadrados*, pois $D(.)$ é definido por uma soma de quadrados.

Para determinar o valor mínimo de uma função de ajuste (ou o seu valor crítico) deve-se derivar parcialmente esta função em relação aos parâmetros β_i's. Então

$$\begin{aligned} \frac{\partial D}{\partial \beta_0} &= 2 \cdot \sum_{i=1}^{n} [y_i - \beta_0 g_0(x_i) - \beta_1 g_1(x) - \cdots - \beta_m g_m(x)] \cdot g_0(x_i) \\ \frac{\partial D}{\partial \beta_1} &= 2 \cdot \sum_{i=1}^{n} [y_i - \beta_0 g_0(x_i) - \beta_1 g_1(x) - \cdots - \beta_m g_m(x)] \cdot g_1(x_i) \\ \frac{\partial D}{\partial \beta_2} &= 2 \cdot \sum_{i=1}^{n} [y_i - \beta_0 g_0(x_i) - \beta_1 g_1(x) - \cdots - \beta_m g_m(x)] \cdot g_2(x_i) \\ \vdots \qquad & \qquad \vdots \\ \frac{\partial D}{\partial \beta_m} &= 2 \cdot \sum_{i=1}^{n} [y_i - \beta_0 g_0(x_i) - \beta_1 g_1(x) - \cdots - \beta_m g_m(x)] \cdot g_m(x_i) \end{aligned}$$

Se $(b_0, b_1, \cdots, b_m)$ for ponto mínimo da função $D(\beta_0, \beta_1, \cdots, \beta_m)$, então

$$\frac{\partial D(b_0, b_1, \cdots, b_m)}{\partial \beta_i} = 0, \quad i = 0, 1, \cdots m$$

e fazendo um rearranjo de termos tem-se:

$$\begin{aligned} \left(\sum_{i=1}^{n} g_0(x_i)^2\right)\beta_0 + \left(\sum_{i=1}^{n} g_0(x_i)g_1(x_i)\right)\beta_1 + \cdots + \left(\sum_{i=1}^{n} g_0(x_i)g_m(x_i)\right)\beta_m &= \sum_{i=0}^{n} y_i g_0(x_i) \\ \left(\sum_{i=1}^{n} g_1(x_i)g_0(x_i)\right)\beta_0 + \left(\sum_{i=1}^{n} g_1(x_i)^2\right)\beta_1 + \cdots + \left(\sum_{i=1}^{n} g_1(x_i)g_m(x_i)\right)\beta_m &= \sum_{i=1}^{n} y_i g_1(x_i) \\ \left(\sum_{i=1}^{n} g_m(x_i)g_0(x_i)\right)\beta_0 + \left(\sum_{i=1}^{n} g_m(x_i)g_1(x_i)\right)\beta_1 + \cdots + \left(\sum_{i=0}^{n} g_m(x_i)^2\right)\beta_m &= \sum_{i=1}^{n} y_i g_m(x_i) \end{aligned}$$

um sistema linear que pode ser solucionado por um método numérico (Gauss, Jordan, etc.). As equações deste sistema são denominadas de equações normais. Observa-se que a matriz dos coeficientes deste sistema é simétrica.

6.3 Ajuste Linear

Um modelo de relacionar duas variáveis é utilizando a equação da reta, caracterizando um comportamento linear do sistema analisado. Se a distribuição dos pontos no diagrama de dispersão ter a aparência de uma reta, então pode-se assumir que:

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= x \\ g_2(x) &= g_3(x) = \cdots = g_m(x) = 0 \end{aligned}$$

que determina o modelo matemático que se ajuste aos pontos do diagrama de dispersão seja a equação da reta, dada por:

$$f(x) = \beta_1 x + \beta_0$$

Utilizando-se o Método dos Mínimos Quadrados

$$D\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left[y_i - \beta_0 - \beta_1 x\right]^2 .$$

Se (b_0, b_1) for ponto mínimo da função $D\left(\beta_0, \beta_1\right)$, então

$$\frac{\partial D\left(b_0, b_1\right)}{\partial \beta_i} = 0, \quad i = 0, 1$$

e fazendo um rearranjo de termos tem-se na notação matricial o seguinte sistema linear:

$$\begin{cases} n\beta_0 & + & \left(\sum_{i=1}^{n} x_i\right)\beta_1 & = & \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\beta_0 & + & \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 & = & \sum_{i=1}^{n} x_i y_i \end{cases}$$

Com solução dada por

$$\beta_1 = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{6.1}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i\right) \cdot \beta_1}{n}$$

Exemplo 6.1

Ajustar os dados da tabela a seguir a uma reta.

i	x_i	y_i
1	1.1	1.8
2	3.6	5.4
3	5.3	4.8
4	6.7	7.3
5	8.0	6.8

Calculando-se os somatórios:

$$\begin{aligned}
\sum_{i=1}^{5} x_i &= 1.1 + 3.6 + 5.3 + 6.7 + 8.0 = 24.7 \\
\sum_{i=1}^{5} x_i^2 &= 1.1^2 + 3.6^2 + 5.3^2 + 6.7^2 + 8.0^2 = 151.15 \\
\sum_{i=1}^{5} y_i &= 1.8 + 5.4 + 4.8 + 7.3 + 6.8 = 26.1 \\
\sum_{i=1}^{5} x_i y_i &= 1.1 \times 1.8 + 3.6 \times 5.4 + 5.3 \times 4.8 + 6.7 \times 7.3 + 8.0 \times 6.8 = 150.17
\end{aligned}$$

Colacando os valores acima nas equações 6.1 obtêm-se

$$\beta_1 = \frac{5 \times 150.17 - 24.7 \times 26.1}{5 \times 151.15 - (24.7)^2} = 0.72896$$

$$\beta_0 = \frac{26.1 - 24.7 \times 0.72896}{5} = 1.6189$$

O script Octave é para traçar o gráfico:

```
x = 0:0.1:10;
X=[1.1;3.6;5.3;6.7;8.0];
Y=[1.8;5.4;4.8;7.3;6.8];
plot (x, 0.72896*x + 1.6189, X, Y, markerstyle = 'o');
xlabel ("x");
ylabel ("y");
title ("Ajuste Linear");
```

O resultado obtido é exibido na figura 6.2.

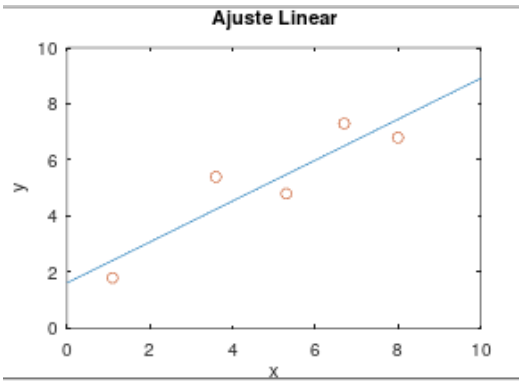


Figura 6.2: Ajuste Linear pelo Método dos Minímos Quadrados.

6.4 Ajuste Polinomial

O ajuste de curva polinomial pode ser realizado quando o diagrama de dispersão não apresenta um comportamento linear. Neste caso, utilizando as seguintes funções $g_i(x)$:

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= x \\ g_2(x) &= x^2 \\ g_3(x) &= x^3 \\ &\vdots \\ g_m(x) &= x^m \end{aligned}$$

Tem-se o seguinte modelo:

$$f(x) = \beta_m x^m + \cdots + \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$$

Sabe-se que polionômios são apropriados para aproximar funções, como por exemplo através de Série de Taylor. Para o ajuste polinomial os parâmetros do modelo são determinados pelo sistema de equações lineares:

$$\begin{cases} n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_2 + \cdots + \left(\sum_{i=1}^{n} x_i^m\right)\beta_m = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^3\right)\beta_2 + \cdots + \left(\sum_{i=1}^{n} x_i^{m+1}\right)\beta_m = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^3\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^4\right)\beta_2 + \cdots + \left(\sum_{i=1}^{n} x_i^{m+2}\right)\beta_m = \sum_{i=1}^{n} x_i^2 y_i \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \quad\quad \vdots \\ \left(\sum_{i=1}^{n} x_i^m\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^{m+1}\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^{m+2}\right)\beta_2 + \cdots + \left(\sum_{i=1}^{n} x_i^{2m}\right)\beta_m = \sum_{i=1}^{n} x_i^m y_i \end{cases}$$

Teorema 6.1 (John von Neumann)

John von Neumann disse a famosa frase: Com quatro parâmetros posso ajustar um elefante, e com cinco posso fazê-lo mexer a tromba. `https://www.johndcook.com/blog/2011/06/21/how-to-fit-an-elephant/`

Exemplo 6.2

Ajustar os dados da tabela a seguir a um polinômio de 2^o grau: $f(x) = \beta_2 x^2 + \beta_1 x + \beta_0$.

i	x_i	y_i
1	0.5	-2.03
2	1.0	-1.35
3	2.5	-0.22
4	3.7	0.59
5	4.7	1.12
6	5.6	1.43

Temos neste exemplo o sistema de equações lineares:

$$\begin{cases} n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_2 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^3\right)\beta_2 = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^3\right)\beta_1 + \left(\sum_{i=1}^{n} x_i^4\right)\beta_2 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

Calculando-se os somatórios:

$$\begin{aligned} \sum_{i=1}^{6} x_i &= 0.5 + 1.0 + 2.5 + 3.5 + 4.7 + 5.6 = 17.8 \\ \sum_{i=1}^{6} x_i^2 &= (-2.03)^2 + (-1.35)^2 + (-0.22)^2 + 0.59^2 + 1.12^2 + 1.43^2 = 73.2 \\ \sum_{i=1}^{6} x_i^3 &= (-2.03)^3 + (-1.35)^3 + (-0.22)^3 + 0.59^3 + 1.12^3 + 1.43^3 = 339.064 \\ \sum_{i=1}^{6} x_i^4 &= (-2.03)^4 + (-1.35)^4 + (-0.22)^4 + 0.59^4 + 1.12^4 + 1.43^4 = 1661.6052 \\ \sum_{i=1}^{6} y_i &= (-2.03) + (-1.35) + (-0.22) + 0.59 + 1.12 + 1.43 = -0.46 \\ \sum_{i=1}^{6} x_i y_i &= (0.5)\cdot(-2.03) + (1.0)\cdot(-1.35) + (2.5)\cdot(-0.22) + (3.5)\cdot(0.59) + \\ &\quad + (4.7)\cdot(1.12) + (5.6)\cdot(1.43) = 12.422 \\ \sum_{i=1}^{6} x_i^2 y_i &= (0.5)^2\cdot(-2.03) + (1.0)^2\cdot(-1.35) + (2.5)^2\cdot(-0.22) + (3.5)^2\cdot(0.59) + \\ &\quad + (4.7)^2\cdot(1.12) + (5.6)^2\cdot(1.43) = 73.5806 \end{aligned}$$

Colocando os valores acima nas equações normais obtêm-se

$$\begin{cases} 6\beta_0 + 17.8\beta_1 + 73.2\beta_2 = -0.46 \\ 17.8\beta_0 + 73.2\beta_1 + 339.064\beta_2 = 12.422 \\ 73.2\beta_0 + 339.064\beta_1 + 1661.6052\beta_2 = 73.5806 \end{cases}$$

a solução usando método de Eliminação:

$\beta_0 = -2.5291, \beta_1 = 1.1587, \beta_2 = -0.0807$

Substituindo estes valores na equação $f(x) = \beta_2 x^2 + \beta_1 x + \beta_0$, obtêm-se

$$f(x) = -0.0807x^2 + 1.1587x - 2.5291$$

O script Octave é para traçar o gráfico:

```
x = linspace (0, 6, 100);
X=[0.5; 1.0; 2.5; 3.5; 4.7; 5.6];
Y=[-2.03 ; -1.35; -0.22 ; 0.59 ; 1.12 ; 1.43];
plot (x,-0.0807*x.^2+ 1.1587*x - 2.5291, X, Y, markerstyle = 'o');
xlabel ("x");
ylabel ("y");
title ("Ajuste Quadrático");
```

O resultado obtido é exibido na figura 6.3.

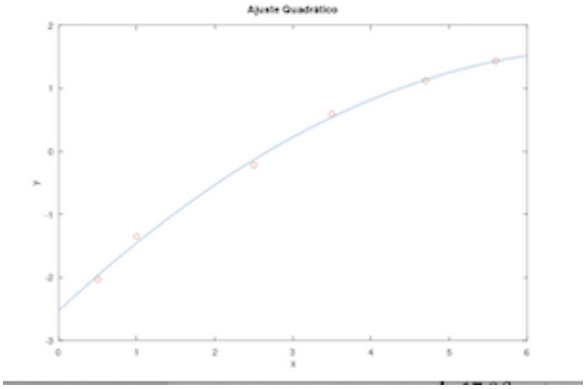


Figura 6.3: Ajuste Quadrático pelo Método dos Minímos Quadrados.

6.5 Ajuste Linear Múltiplo

O ajuste de curva linear múltiplo pode ser realizado quando uma variável dependente (resposta) é uma função com duas ou mais variáveis independentes (explicativas) onde cada variável explicativa tem uma relação linear com a de resposta. É equivalente a fazer um ajuste linear num espaço n-dimensional, ao invés de 2-dimensões. Neste caso, utilizando as seguintes funções $g_i(x)$:

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= x_1 \\ g_2(x) &= x_2 \\ g_3(x) &= x_3 \\ &\vdots \\ g_m(x) &= x_m \end{aligned}$$

Tem-se o seguinte modelo:

$$f(x_1, x_2, \cdots, x_m) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x^m$$

Para o ajuste de curva linear múltiplo os parâmetros do modelo são determinados pelo sistema de equações lineares:

$$\begin{cases} n\beta_0 + \left[\sum_{i=1}^{n}(x_1)_i\right]\beta_1 + \left[\sum_{i=1}^{n}(x_2)_i\right]\beta_2 + \cdots + \left[\sum_{i=1}^{n}(x_m)_i\right]\beta_m = \sum_{i=1}^{n} y_i \\ \left[\sum_{i=1}^{n}(x_1)_i\right]\beta_0 + \left[\sum_{i=1}^{n}(x_1)_i^2\right]\beta_1 + \left[\sum_{i=1}^{n}(x_1)_i(x_2)_i\right]\beta_2 + \cdots + \left[\sum_{i=1}^{n}(x_1)_i(x_m)_i\right]\beta_m = \sum_{i=1}^{n}(x_1)_i y_i \\ \left[\sum_{i=1}^{n}(x_2)_i\right]\beta_0 + \left[\sum_{i=1}^{n}(x_2)_i(x_1)_i\right]\beta_1 + \left[\sum_{i=1}^{n}(x_2)_i^2\right]\beta_2 + \cdots + \left[\sum_{i=1}^{n}(x_2)_i(x_m)_i\right]\beta_m = \sum_{i=1}^{n}(x_2)_i y_i \\ \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ \left[\sum_{i=1}^{n}(x_m)_i\right]\beta_0 + \left[\sum_{i=1}^{n}(x_m)_i(x_1)_i\right]\beta_1 + \left[\sum_{i=1}^{n}(x_m)_i^2\right]\beta_2 + \cdots + \left[\sum_{i=1}^{n}(x_m)_i^2\right]\beta_m = \sum_{i=1}^{n}(x_m)_i y_i \end{cases}$$

Exemplo 6.3

A tabela a seguir foi criada usando a equação:

$$y(x_1, x_2) = 5x_1 - 3x_2$$

i	$(x_1)_i$	$(x_2)_i$	$y_i(x_1, x_2)$
1	1	1	12
2	2	1	17
3	1	3	6
4	4	7	9
5	7	2	39
6	9	10	25

Use ajuste linear múltiplo para modelar estes dados.

Temos neste exemplo o sistema de equações lineares:

$$\begin{cases} n\beta_0 + \left[\sum_{i=1}^{n}(x_1)_i\right]\beta_1 + \left[\sum_{i=1}^{n}(x_2)_i\right]\beta_2 = \sum_{i=1}^{n} y_i \\ \left[\sum_{i=1}^{n}(x_1)_i\right]\beta_0 + \left[\sum_{i=1}^{n}(x_1)_i^2\right]\beta_1 + \left[\sum_{i=1}^{n}(x_1)_i(x_2)_i\right]\beta_2 = \sum_{i=1}^{n}(x_1)_i y_i \\ \left[\sum_{i=1}^{n}(x_2)_i\right]\beta_0 + \left[\sum_{i=1}^{n}(x_2)_i(x_1)_i\right]\beta_1 + \left[\sum_{i=1}^{n}(x_2)_i^2\right]\beta_2 = \sum_{i=1}^{n}(x_2)_i y_i \end{cases}$$

Calculando-se os somatórios:

$$\begin{aligned}
\sum_{i=1}^{6}(x_1)_i &= 1+2+1+4+7+9 = 24\\
\sum_{i=1}^{6}(x_2)_i &= 1+1+3+7+2+10 = 24\\
\sum_{i=1}^{6}(x_1)_i^2 &= 1^2+2^2+1^2+4^2+7^2+9^2 = 152\\
\sum_{i=1}^{6}(x_2)_i^2 &= 1^2+1^2+3^2+7^2+2^2+10^2 = 164\\
\sum_{i=1}^{6}(x_1)_i(x_2)_i &= 1\cdot 1+2\cdot 1+1\cdot 3+4\cdot 7+7\cdot 2+9\cdot 10 = 138\\
\sum_{i=1}^{6} y_i &= 12+17+6+9+39+25 = 108\\
\sum_{i=1}^{6}(x_1)_i y_i &= 1*12+2*17+1*6+4*9+7*39+9*25 = 586\\
\sum_{i=1}^{6}(x_2)_i y_i &= 1*12+1*17+3*6+7*9+2*39+10*25 = 438
\end{aligned}$$

Colacando os valores acima nas equações normais obtêm-se

$$\begin{cases}
6\beta_0 + 24\beta_1 + 24\beta_2 = 108\\
24\beta_0 + 152\beta_1 + 138\beta_2 = 586\\
24\beta_0 + 138\beta_1 + 164\beta_2 = 438
\end{cases}$$

E a solução usando método de Eliminação:

$$\beta_0 = 10, \beta_1 = 5, \beta_2 = -3$$

é consistente com a equação original da qual os dados foram derivados.

6.6 Linearização de Relações não Lineares

A método dos mínimos quadrados fornece uma técnica poderosa para ajustar a melhor linha aos dados. Existem muitas situações na ciência e engenharia que mostram a relação entre as quantidades que estão sendo consideradas não é linear. Existem vários exemplos de funções não lineares usadas para ajuste de curvas. Alguns deles estão nas Tabelas 6.1 e 6.2.

Técnicas de ajuste dos mínimos não linear estão disponíveis para ajustar essas equações das Tabelas 6.1 e 6.2 aos dados diretamente. Uma alternativa mais simples é usar manipulações analíticas para transformar as equações em uma forma linear que pode ser usada para ajustar as equações aos dados.

N^o	Equação	Forma Linear	$\overline{y} = b_1\overline{x} + b_0$
1	$y = \beta_0 x^{\beta_1}$	$ln(y) = \beta_1 ln(x) + ln(\beta_0)$	$\overline{y} = ln(y), \overline{x} = ln(x)$ $b_1 = \beta_1, b_0 = ln(\beta_0)$
2	$y = \beta_0 e^{\beta_1 x}$	$ln(y) = \beta_1 x + ln(\beta_0)$	$\overline{y} = ln(y), \overline{x} = x$ $b_1 = \beta_1, b_0 = ln(\beta_0)$
3	$y = \beta_0 10^{\beta_1 x}$	$log(y) = \beta_1 x + log(\beta_0)$	$\overline{y} = log(y), \overline{x} = x$ $b_1 = \beta_1, b_0 = log(\beta_0)$
4	$y = \frac{1}{\beta_1 x + \beta_0}$	$\frac{1}{y} = \beta_1 x + \beta_0$	$\overline{y} = \frac{1}{y}, \overline{x} = x$ $b_1 = \beta_1, b_0 = \beta_0$
5	$y = \frac{\beta_1 x}{\beta_0 + x}$	$\frac{1}{y} = \frac{\beta_0}{\beta_1 x} + \frac{1}{\beta_1}$	$\overline{y} = \frac{1}{y}, \overline{x} = \frac{1}{x}$ $b_1 = \frac{\beta_0}{\beta_1}, b_0 = \frac{1}{\beta_1}$

Tabela 6.1: Transformações lineares para Ajuste Linear

N^o	Equação	Forma Linear Múltipla
1	$y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}$	$ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
2	$y = \beta_0 \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdot \cdots \cdot x_n^{\beta_n}$	$ln(y) = ln(\beta_0) + \beta_1 ln(x_1) + \beta_2 ln(x_2) + \cdots + \beta_n ln(x_n)$
3	$y = \frac{1}{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n}$	$\frac{1}{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$
3	$y = \frac{1}{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n}$	$\frac{1}{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$

Tabela 6.2: Transformações lineares para Ajuste Linear Múltiplo

Exemplo 6.4

Ajustar os dados da tabela a seguir a equação $y = \beta_0 e^{\beta_1}$.

i	x_i	y_i
1	0.1	2.9
2	1.6	4.4
3	3.4	6.0
4	4.5	9.9
5	5.0	10.8

Fazendo: $\begin{cases} \overline{y} = ln(y), \overline{x} = x \\ b_1 = \beta_1, b_0 = ln(\beta_0) \end{cases}$, então tem-se a tabela seguinte:

i	x_i	y_i
1	0.1	1.06
2	1.6	1.48
3	3.4	1.79
4	4.5	2.29
5	5.0	2.38

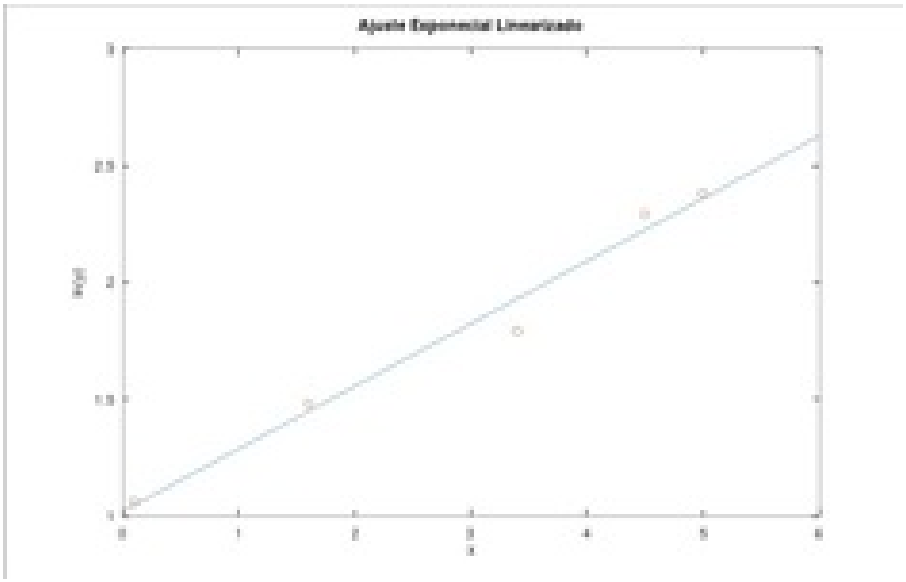


Figura 6.4: Ajuste Exponecial pelo Método dos Mínimos Quadrados Linearizado.

Calculando-se os somatórios:

$$\sum_{i=1}^{5} x_i = 0.1 + 1.6 + 3.4 + 4.5 + 5.0 = 14.6$$

$$\sum_{i=1}^{5} x_i^2 = 0.1^2 + 1.6^2 + 3.4^2 + 4.5^2 + 5.0^2 = 59.38$$

$$\sum_{i=1}^{5} y_i = 1.06 + 1.48 + 1.79 + 2.29 + 2.38 = 9$$

$$\sum_{i=1}^{5} x_i y_i = 0.1 * 1.06 + 1.6 * 1.48 + 3.4 * 1.79 + 4.5 * 2.29 + 5 * 2.38 = 30.765$$

Colocando os valores acima nas equações 6.1 obtêm-se

$$b_1 = \frac{5 \times 30.765 - 14.6 \times 9}{5 \times 59.38 - (14.6)^2} = 0.26779$$

$$b_0 = \frac{9 - 14.6 \times 0.26779}{5} = 1.01804$$

O *script* Octave é para traçar o gráfico:

```
x = 0:0.1:6;
X=[0.1;1.6;3.4;4.5;5.0];
Y=[1.06;1.48;1.79;2.29;2.38];
plot (x, 0.26779*x + 1.01804, X, Y, markerstyle = 'o');
xlabel ("x");
ylabel ("ln(y)");
title ("Ajuste Exponecial Linearizado");
```

O resultado obtido é exibido na figura 6.4.

Como $b_0 = ln(\beta_0)$ e $b_1 = \beta_1$, tem-se como estimativas dos parâmetros originais: $\beta_0 = 2.76776$ e $\beta_1 = 0.26779$ que resulta na função $y(x) = 2.76776 \cdot e^{0.26779x}$.

O script Octave é para traçar o gráfico:

```
x = 0:0.1:6;
X=[0.1;1.6;3.4;4.5;5.0];
```

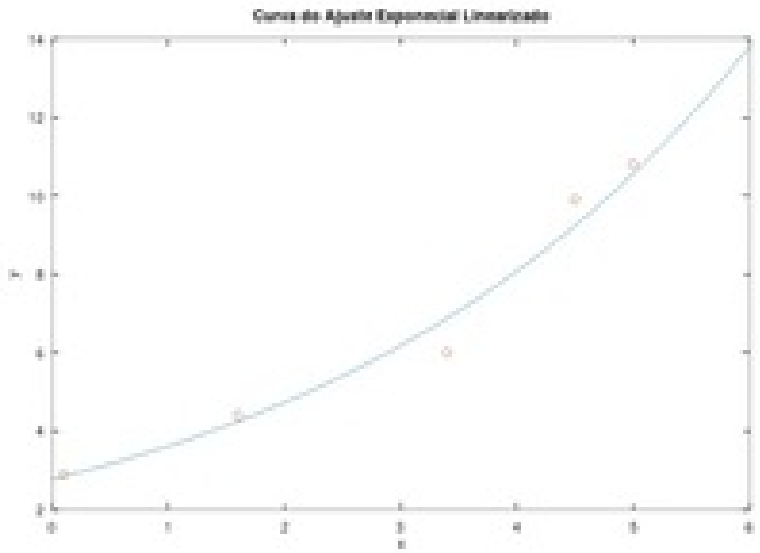


Figura 6.5: Cuva Original do Ajuste Exponecial pelo Método dos Minímos Quadrados Linearizado.

```
Y=[2.9;4.4;6.0;9.9;10.8];
plot (x, 2.76776*exp(0.26779*x), X, Y, markerstyle = 'o');
xlabel ("x");
ylabel ("y");
title ("Curva do Ajuste Exponecial Linearizado");
```

O resultado obtido é exibido na figura 6.5.

Apêndice

Os *scripts* Octave são para traçar o gráfico 6.1: `pkg load statistics`

```
x = linspace (2, 8, 100);
rand('seed',4)
X=unifrnd(2,8,[1,100]);
N=normrnd(0,1,[1,100]);
Y=X.^2 - 10*X + 30 + N;
plot (x,x.^2 - 10*x + 30, 'linewidth',5, X, Y, markerstyle = 'o','linewidth',5);
h=get(gcf, "currentaxes");
set(h, "fontsize", 12, "linewidth", 2);
xlabel ("x");
ylabel ("y");
title ("Modelo de Regressão");

p = polyfit (X, Y, 1);
plot (x,p(1)*x + p(2), 'linewidth',5, X, Y, markerstyle = 'o','linewidth',5);
h=get(gcf, "currentaxes");
set(h, "fontsize", 12, "linewidth", 2);
xlabel ("x");
ylabel ("y");
title ("Sub Ajuste");

p = polyfit (X, Y, 2)
y = polyval (p, x)
plot (x,y, 'linewidth',5, X, Y, markerstyle = 'o','linewidth',5);
h=get(gcf, "currentaxes");
set(h, "fontsize", 12, "linewidth", 2);
xlabel ("x");
ylabel ("y");
title ("Ajuste Ideal");

p = polyfit (X, Y, 20)
y = polyval (p, x)
plot (x,y,'linewidth',5, X, Y, markerstyle = 'o','linewidth',5);
h=get(gcf, "currentaxes");
set(h, "fontsize", 12, "linewidth", 2);
xlabel ("x");
ylabel ("y");
title ("Sobre Ajuste");
```

Capítulo Execícios

1. Ajustar os pontos da tabela seguinte a uma reta.

x_i	1	2.5	4.2	5.1	6.5	8.4
y_i	4.4	3.9	3.1	1.7	0.1	-0.5

2. Dada tabela abaixo: https://doi.org/10.1590/S1806-11172009000300010

Tabela - Altura $h(m)$ do corpo em queda livre e intervalo de tempo $t(s)$.

$t(s)$	$h(m)$	$t(s)$	$h(m)$
0.033	0.053	0.300	0.559
0.067	0.083	0.333	0.677
0.100	0.123	0.367	0.789
0.133	0.165	0.400	0.924
0. 167	0.224	0.433	1.083
0.200	0.283	0.467	1.235
0.233	0.365	0.500	1.400
0.267	0.459	0.533	1.577

a) Ajuste uma parábola $h(t) = a + b \cdot t$ utilizando os pontos:
$\{(0.1, 0.123), (0.2, 0.283), (0.3, 0.559), (0.4, 0.924), (0.5, 1.400)\}$;
b) Calcule os valores de $h(t)$ obtido acima para estimar a altura $\widetilde{h}(t)$ nos tempos indicados e calcular o erro para preencher a tabela abaixo:

$t(s)$	$h(m)$	$\widetilde{h}(t)$	$Erro = \left\lvert h(m) - \widetilde{h}(t) \right\rvert$
0.033	0.053		
0.067	0.083		
0.133	0.165		
0.167	0.224		
0.233	0.365		
0.267	0.459		
0.333	0.677		
0.367	0.789		
0.433	1.083		
0.467	1.235		
0.533	1.577		

3. Segue 20 coleta dados sobre a porcentagem de pessoas em cada cidade que fumam, a porcentagem de pessoas em cada cidade que vão de bicicleta para o trabalho e a porcentagem de pessoas em cada cidade que têm doenças cardíacas.
(adaptado de: https://www.scribbr.com/statistics/multiple-linear-regression/)

Tabela - Taxas de cardiopatias (% da população) em função do deslocamento de bicicleta para o trabalho e tabagismo

Cardiopatias (%)	*Deslocamento de bicicleta* (%)	*Tabagismo* (%)
30.80	10.90	11.77
65.13	2.22	2.85
1.96	17.59	17.18
44.80	2.80	6.82
69.43	15.97	4.06
54.40	29.33	9.55
49.06	9.06	7.62
4.78	12.84	15.85
65.73	11.99	3.07
35.26	23.28	12.10
51.83	14.44	6.43
52.94	25.07	8.61
48.77	11.02	6.72
26.17	6.64	10.60
10.55	5.99	14.08
47.16	14.10	8.74
61.68	16.84	5.44
33.94	5.76	9.16
39.70	12.66	9.75
63.12	22.92	5.86

a) Faça o ajuste linear múltiplo considerando as duas variáveis independentes,$x_1 =$ *deslocamento de bicicleta para o trabalho*, $x_2 =$ *tabagismo* e como variável dependente $y =$ *cardiopatias*.

4. AJUSTE DE CURVAS EXPONENCIAL

As medidas das taxas na qual a radiação de partículas emitidas por ^{55}Fe são detectadas quando um contador Geiger é blindado com uma ou mais folhas de alumínio com aparece na tabela abaixo:

Expessura de Al (cm)	Taxa de Contagem (cotagem/min)
0.00162	1850
0.00324	1250
0.00486	800
0.00648	450
0.00810	310
0.00972	165

A absorção da radiação para uma dada expessura de material pode ser modelado pela relação exponencial

$$R(x) = R_0 e^{\beta x}$$

onde $R(x)$ é a taxa de contagem da radiação de particulas, R_0 é a taxa de contagem sem a presença da blindagem, x é a espessura do material, e β é uma constante que descreve quão rapidamente a taxa de contagem decresce com o crescimento da blindagem.

a) Determine R_0 e β;
b) Verifique que $\beta = 290\, cm^{-1}$.

Capítulo - Equação Diferencial Ordinária

Uma equação diferencial ordinária (EDO), pode ser considerada como uma igualdade diferencial especificando a relação entre uma variável dependente y, e uma variável independente x. A ordem da EDO é a ordem da maior derivada de y em relação a x aparecendo nela.

$\frac{dy}{dx} + xy^3 = 0$ é uma EDO de primeira ordem,

$\frac{d^2y}{dx^2} - ky = 0$ é uma EDO de segunda ordem.

Uma EDO é linear se todos as potências de y e suas derivadas que aparecem na EDO são inteiros não negativos que não excedem a unidade. A equação geral para uma EDO linear de ordem n é

$$a_n(x)\frac{d^n y}{dx^n} + a_{n-1}(x)\frac{d^{n-1}y}{dx^{n-1}} + ... + a_1(x)\frac{dy}{dx} + a_0(x)y = f(x). \tag{7.1}$$

Diz-se que a EDO linear da Eq. 7.1 é homogênea se $f(x) \equiv 0$.

Faz-se a distinção entre o problema de valor inicial (PVI) e o problema de valor no contorno (PVC). Essa classificação baseia-se na especificação de dados complementares que nos permitem chegar a uma solução única da EDO. Para o problema de valor inicial, todos os dados (condições iniciais) são especificados em um ponto, enquanto para o problema de valor no contorno os dados são dados como condições no contorno, ou seja, como condições de contorno. Por exemplo, considere a EDO de primeira ordem

$$\frac{dy}{dx} = y.$$

A solução é uma família de curvas dadas por $y = a \cdot exp(x)$. Para determinar a solução única, é preciso fornecer o valor de y para algum valor de x. Suponha que tenha sido dado pela condição $y(0) = 1$. Encontra-se $a = 1$, daí a solução para o PVI

$$\frac{dy}{dx} = y, \quad y(0) = 1 \tag{7.2}$$

é dada unicamente por

$$y = \exp(x).$$

Na declaração do PVI acima, chama-se a condição $y(0) = 1$ a condição inicial. A origem desta terminologia é por causa da dependência da variável x com o tempo t em muitos problemas físicos. Por exemplo, de acordo com a mecânica clássica, o movimento dos objetos é regido pela segunda ordem PVI decorrente da segunda lei de Newton

$$\begin{aligned} m\frac{d^2x}{dt^2} &= \sum_i F_i \\ \left(\frac{dx}{dt}\right)_{(t=0)} &= v_0 \\ x(t=0) &= x_0, \end{aligned} \tag{7.3}$$

onde x é a posição do objeto e t o tempo. Observe que o PVI acima tem 2 (duas) condições iniciais. Em geral, você precisa de tantas condições iniciais quanto a ordem do PVI para obter uma solução única.

Observe que pode-se escrever o PVI de segunda ordem acima (Eq.7.3) como dois EDOs de primeira ordem mais uma condição inicial para cada EDO. Para conseguir isso, defini-se $v \equiv dx/dt$. Então o sistema de EDOs que representam o mesmo PVI é

$$\begin{aligned} m\frac{dv}{dt} &= \sum_i F_i \\ \frac{dx}{dt} &= v \\ v(t=0) &= v_0 \\ x(t=0) &= x_0, \end{aligned} \tag{7.4}$$

Como exemplo de um problema de valor de contorno (PVC), considere a seguinte equação para o perfil de temperatura de estado estacionário em uma vara unidimensional de comprimento l, mantida a uma temperatura constante em uma extremidade e isolada na outra, com uma fonte de calor/sumidouro presente dentro da haste.

$$\begin{aligned} &\frac{\partial^2 T}{\partial x^2} = f(T,x) \\ &T(x=0) = T_0 \\ &\left(\frac{\partial T}{\partial x}\right)_{(x=l)} = 0. \end{aligned} \tag{7.5}$$

Observe que os dados são fornecidos aos dois pontos extremos da haste. Estes tipos de condições são chamadas condições de contorno e o PVC dado descreve o perfil de temperatura dentro da haste sujeito às condições de limite especificadas acima.

O EDO que aparece no PVC acima (Eq. 7.5) é muito semelhante ao EDO de segunda ordem que aparece no PVI na (Eq. 7.3). No entanto, do ponto de vista matemático, esses problemas são muito diferentes, a diferença decorrente principalmente da especificação das condições iniciais/de fronteira. Assim, os algoritmos para a solução numérica de PVIs e PVCs também são significativamente diferentes.

7.1 Solução numérica de problemas de valores iniciais

Alguns dos conceitos-chave associados à solução numérica dos PVIs são o *erro de truncamento*, *erro local*, a *ordem* e a *estabilidade* do método numérico. Devemos também ser capazes de distinguir técnicas explícitas das implícitas. A seguir, esses conceitos serão introduzidos através de exemplos simples.

7.1.1 Método de Euler Explicíto

Será vista a solução numérica do PVI, escrito na forma explícita,

$$\frac{dy}{dx} = f(x, y(x)), \quad y(x=0) = y_0. \tag{7.6}$$

Em particular, se $f(y,t) \equiv g(y)$, o PVI acima com $g(y) = ky$ onde k é uma constante, o PVI é linear. Assumimos que existe uma solução única e denotamos essa solução por $y^e(t)$. Assim, a partir de agora, y(t) refere-se à solução calculada numericamente, que na melhor das hipóteses é apenas uma aproximação a $y^e(t)$.

Considere que os seguintes objetos sejam dados: alguma EDO explícitada na forma (Eq. 7.2), uma condição inicial (x_0, y_0) e um domínio de solução desejado $[x_0, x_n]$. Uma abordagem de solução simples é discretizar o domínio da solução $[x_0, x_n]$ em $n+1$ pontos,

$$x_0 < x_1 < \cdots < x_n$$

e aproximar as derivadas $y'(x_{i-1}) = \left(\frac{dy}{dx}\right)_{(x_{i-1})}$ por diferenças finitas $\frac{y_i - y_{i-1}}{x_i - x_{i-1}}$ para $i = 1, \cdots, n$, onde y_i é uma aproximação para o desconhecido $y(x_i)$. Com essa aproximação, a equação (7.6) avaliada no pontos $x_{i?1}$ torna-se

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} = f(x_{i-1}, y_{i-1})$$

A qual pode ser resolvida para y_i:

$$y_i = y_{i-1} + f(x_{i-1}, y_{i-1})(x_i - x_{i-1}). \tag{7.7}$$

Exemplo 7.1

Considere o problema de valor inicial

$$y'(x) = 2y(x), y(0) = 0$$

cuja solução é $y(x) = e^{2x}$. O método de Euler aplicado a este problema produz o esquema:

$$y_i = y_{i-1} + 2 \cdot h \cdot y_{i-1} = (1 + 2h) \cdot y_{i-1} \tag{7.8}$$

Suponha que queremos calcular o valor aproximado de $y(1)$ com $h = x_i - x_{i-1} = 0.2$. As aproximações para a solução nos pontos da malha (valores de x_i) usando o método de Euler são:

$$\begin{aligned}
y(0) &= y_1 = 1 \\
y(0.2) &= y_2 = (1 + 2h)y_1 = 1.4y_1 = 1.4 \\
y(0.4) &= y_3 = (1 + 2h)y_2 = 1.4y_2 = 1.96 \\
y(0.6) &= y_4 = (1 + 2h)y_3 = 1.4y_3 = 2.744 \\
y(0.8) &= y_5 = (1 + 2h)y_4 = 1.4y_4 = 3.8416 \\
y(1.0) &= y_6 = (1 + 2h)y_5 = 1.4y_5 = 5.37824
\end{aligned}$$

Essa aproximação é bem inexata quando comparamos com a solução do problema em $y(1) = e^2 = 7.38906$. Escolhendo um passo menor, obteremos uma aproximação melhor. Veja tabela abaixo com valores obtidos com diferentes valores de passo .

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}
$y(1)$	6.7275	7.2446	7.3443	7.3876

Nete caso é possível mostrar que quando $h \to 0+$ a solução aproximada via método de Euler converge para a solução exata .

A solução da relação de recorrência (Eq. 7.8) é dada por

$$y_k = (1 + 2h)^{k-1}$$

Como $x_k = (k - 1) \cdot h$ e queremos a solução em $x = 2$, a solução aproximada pelo método de Euler com passo h em é dada por:

$$y_k = (1 + 2h)^{k-1} = (1 + 2h)^{\frac{2}{h}}$$

Aplicando o limite$h \to 0+$, tem-se:

$$\lim_{h \to 0+} (1 + 2h)^{\frac{2}{h}} = e^2$$

7.1.2 Método de Euler Implicíto

Uma equação diferencial ordinária de primeira ordem é uma equação da forma implícita

$$F(x, y, y') = 0 \tag{7.9}$$

com uma função $F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. Uma solução é uma função diferenciável $y : \mathbb{R} \to \mathbb{R}^n$ com a propriedade

$$\forall x \in \mathbb{R} : F(y(x), y'(x), x) = 0$$

Difinições são válidas, devidamente ajustadas, se F for definida num subconjunto de $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ e/ou se a só um subconjunto na solução y é de interesse.

Considere uma condição inicial (x_0, y_0) ,um domínio de solução $[x_0, x_n]$ e uma discretização $\{x\}_{i=0}^n$ do domínio ser dado. O método de Euler aproxima $y'(x)$ por $\dfrac{y_i - y_{i-1}}{x_i - x_{i-1}}$ e usa a EDO nos pontos $x_0, x_1, \cdots, x_n$, o que conduz as equações

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} = f(x_i, y_i), \quad i = 1, \cdots, n$$

Não existe geralmente uma solução explicita para y_i. Agora é preciso resolver em cada passo um sistema de equações – potencialmente nonlinear para encontrá-las. Embora isso parece ser muito mais complicado do que o método explícito, ele tem muitas vezes propriedades de estabilidade superiores.

Exemplo 7.2

Digamos que se queira resolver

$$\frac{dy}{dx} = y \cdot sen(y)$$

Euler implícto resulta em

$$y_n = y_{n-1} + h \cdot y_n \cdot sen(y_n)$$

ou

$$y_n(1 - h \cdot sen(y_n) = y_{n-1}$$

o qual é claramente não linear.

Este método não exige que a EDO seja dada de forma explícita (Eq. 7.6). Ele funciona igualmente para um EDO implícito (Eq. 7.9). Então o sistema de equação a ser resolvido em cada etapa é

$$F\left(x_i, y_i, \frac{y_i - y_{i-1}}{x_i - x_{i-1}} = y'\right) = 0 \tag{7.10}$$

7.2 Outros Métodos

Há muitos outros métodos que não cobrimos neste curso. Duas classes comuns de métodos são (ambos existem em variantes implícitas e explícitas e hibridas):

- Métodos de Runge-Kutta: como o método de Euler, eles são métodos de etapa única, pois ao computar a solução y_i ao longo de um intervalo de discretização $[x_{i-1}, x_i]$, eles só usam as informações do valor da função da última etapa (x_{i-1}, y_{i-1}) como uma condição inicial. Cada etapa é calculada usando vários estágios com o objetivo de aumentar a ordem de convergência. Cada etapa envolve uma avaliação de f ou - no caso de métodos implícitos - uma equação a ser resolvida.
- Métodos de várias etapas: usa informações (x_j, y_j, y'_j), a partir de vários passos anteriores $j \leq i - 1$ quando da computação de (y_i, y'_i), com o objetivo de ganhar eficiência. Versões implícitas de métodos multi-passo (lineares) têm propriedades de estabilidade piores do que certos métodos de passo único para ordens de convergência maiores do que 2.

7.2.1 Métodos Runge-Kutta

Na análise numérica, os métodos Runge-Kutta são uma família de métodos iterativos implícitos e explícitos, que incluem o Método Euler, usada nas soluções aproximadas de equações diferenciais ordinárias.

Lista de métodos Runge-Kutta:

https://en.wikipedia.org/wiki/List_of_Runge%E2%80%93Kutta_methods

1. Métodos explícitos
   (a). Euler direto
   (b). Método de ponto médio explícito
   (c). Método de Heun
   (d). Método de Ralston
   (e). Método genérico de segunda ordem
   (f). Método de terceira ordem de Kutta
   (g). Método genérico de terceira ordem
   (h). Método de terceira ordem de Heun
   (i). Método de terceira ordem de Ralston
   (j). Estabilidade forte de terceira ordem Preservando Runge-Kutta (SSPRK3)

(k). Método clássico de quarta ordem
(l). Método de quarta ordem de Ralston
(m). Método de quarta ordem regra-3/8

2. Métodos embutidos
   (a). Heun-Euler
   (b). Fehlberg RK1 e RK2 (ordem)
   (c). Bogacki-Shampine
   (d). Fehlberg
   (e). Cash-Karp
   (f). Dormand-Prince
3. Métodos implícitos
   (a). Euler retroativo
   (b). Ponto médio implícito
   (c). Método Crank-Nicolson
   (d). Métodos Gauss-Legendre
   (e). Métodos diagonalmente implícitos de Runge-Kutta
   (f). Métodos Lobatto
      I. Métodos Lobatto IIIA
      II. Métodos Lobatto IIIB
      III. Métodos Lobatto IIIC
      IV. Métodos Lobatto IIIC*
      V. Métodos de Lobatto generalizados
   (g). Métodos Radau
      I. Métodos Radau IA
      II. Métodos Radau IIA

7.2.2 Método RK-2

No método de Euler explícito direto, usa-se as informações sobre a inclinação ou a derivada de y no intervalo x dado para extrapolar a solução para o próximo intervalo de x. O erro de truncamento local (ETL) para o método é $O(h^2)$, resultando em uma técnica numérica de primeira ordem. Em geral, um método com $O(h^{k+1})$ ETL é considerado de ordem k. Os métodos de Runge-Kutta são uma classe de métodos que usa criteriosamente as informações da 'inclinação' em mais de um ponto para extrapolar a solução para o intervalo de tempo futuro.

O método de segunda ordem Runge-Kutta (RK2), onde o ETL é $O(h^3)$, para aproximar a solução do problema de valor inicial $y'(x) = f(x, y); y(x_0) = y_0$, avalia o integrando,$f(x, y)$,duas vezes por passo. Para o passo n,

$$\begin{aligned} k_1 &= hf(x_{n-1}, y_{n-1}) \\ k_2 &= hf(x_{n-1} + h, y_{n-1} + k_1) \\ y_n &= y_{n-1} + \frac{(k_1 + k_2)}{2}, \qquad \text{(RK2).} \\ &\text{e} \\ x_n &= x_0 + n \cdot h \end{aligned} \tag{7.11}$$

Exemplo 7.3

Considere o problema de valor inicial

$$y'(x) = 2y(x), y(0) = 0$$

cuja solução é $y(x) = e^{2x}$. O método RK2 será aplicado a este problema.

Suponha que queremos calcular o valor aproximado de $y(1)$ com $h = x_i - x_{i-1} = 0.2$. As aproximações para a

solução nos pontos da malha (valores de x_i) usando o RK2 são:

$$
\begin{aligned}
&k_1 = hf(x_0, y_0) = (0.2)f(0,1) = (0.2)\cdot(2) = 0.4\\
&k_2 = hf(x_0 + h, y_0 + k_1) = (0.2)f(0.2, 1.4) = (0.2)\cdot(2.8) = 0.56\\
&y_1 = y_0 + \frac{k_1 + k_2}{2} = 1 + 0.48 = 1.48\\
&y(0.2) = 1.48
\end{aligned}
$$

$$
\begin{aligned}
&k_1 = hf(x_1, y_1) = (0.2)f(0.2, 1.48) = (0.2)\cdot(2.96) = 0.592\\
&k_2 = hf(x_1 + h, y_1 + k_1) = (0.2)f(0.4, 2.072) = (0.2)\cdot(4.144) = 0.8288\\
&y_2 = y_1 + \frac{k_1 + k_2}{2} = 1.48 + 0.7104 = 2.1904\\
&y(0.4) = 1.48
\end{aligned}
$$

$$
\begin{aligned}
&k_1 = hf(x_2, y_2) = (0.2)f(0.4, 2.1904) = (0.2)\cdot(4.3808) = 0.8762\\
&k_2 = hf(x_2 + h, y_2 + k_1) = (0.2)f(0.6, 3.0666) = (0.2)\cdot(6.1331) = 1.2266\\
&y_3 = y_2 + \frac{k_1 + k_2}{2} = 2.1904 + 1.0514 = 3.2418\\
&y(0.6) = 3.2418
\end{aligned}
$$

$$
\begin{aligned}
&k_1 = hf(x_3, y_3) = (0.2)f(0.6, 3.2418) = (0.2)?(6.4836) = 1.2967\\
&k_2 = hf(x_3 + h, y_3 + k_1) = (0.2)f(0.7, 3.8902) = (0.2)\cdot(7.7803) = 1.5561\\
&y_4 = y_3 + \frac{k_1 + k_2}{2} = 3.2418 + 1.5561 = 4.7979\\
&y(0.8) = 4.7979
\end{aligned}
$$

$$
\begin{aligned}
&k_1 = hf(x_4, y_4) = (0.2)f(0.8, 4.7979) = (0.2)\cdot(9.5957) = 1.9191\\
&k_2 = hf(x_4 + h, y_4 + k_1) = (0.2)f(0.9, 5.7574) = (0.2)\cdot(11.5148) = 2.303\\
&y_5 = y_5 + \frac{k_1 + k_2}{2} = 4.7979 + 2.303 = 7.1008\\
&y(1) = 7.1008
\end{aligned}
$$

Essa aproximação é inexata quando comparamos com a solução do problema em $y(1) = e^2 = 7.38906$, mas já é melhor que a obtida com o método de Euler com $h = 2$. Escolhendo um passo menor, obteremos uma aproximação melhor. Veja tabela abaixo com valores obtidos com diferentes valores de passo .

h	10^{-1}	10^{-2}	10^{-3}
$y(1)$	7.3046	7.3881	7.3890

7.2.3 Método RK-3

O método de terceira ordem Runge-Kutta (RK3), onde o ELT é $O(h^4)$, para aproximar a solução do problema de valor inicial $y'(x) = f(x,y); y(x_0) = y_0$, avalia o integrando,$f(x,y)$, três vezes por passo. Para o passo n,

$$
\begin{aligned}
&k_1 = hf(x_{n-1}, y_{n-1})\\
&k_2 = hf(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{k_1}{2})\\
&k_3 = hf(x_{n-1} + h, y_{n-1} + 2k_2 - k_1)\\
&y_n = y_{n-1} + \frac{1}{6}(k_1 + 4k_2 + k_3), \qquad \text{(RK3)}\\
&\text{e}\\
&x_n = x_0 + n\cdot h
\end{aligned}
\tag{7.12}
$$

Exemplo 7.4

Considere o problema de valor inicial

$$y'(x) = 2y(x), y(0) = 0$$

cuja solução é $y(x) = e^{2x}$. O método RK3 será aplicado a este problema.

Suponha que queremos calcular o valor aproximado de $y(1)$ com $h = x_i - x_{i-1} = 0.2$. As aproximações para a solução nos pontos da malha (valores de x_i) usando o RK3 são:

$$\begin{aligned}
&k_1 = hf(x_0, y_0) = (0.2)f(0, 1) = (0.2) \cdot (2) = 0.4\\
&k_2 = hf(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}) = (0.2)f(0.1, 1.2) = (0.2) \cdot (2.4) = 0.48\\
&k_3 = hf(x_0 + h, y_0 + 2k_2 - k_1) = (0.2)f(0.2, 1.56) = (0.2) \cdot (3.12) = 0.624\\
&y_1 = y_0 + \frac{1}{6}(k_1 + 4k_2 + k_3) = 1 + \frac{1}{6}[0.4 + 4(0.48) + (0.624)] = 1.4907\\
&y(0.2) = 1.4907
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_1, y_1) = (0.2)f(0.2, 1.4907) = (0.2) \cdot (2.9813) = 0.5963\\
&k_2 = hf(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}) = (0.2)f(0.3, 1.7888) = (0.2) \cdot (3.5776) = 0.7155\\
&k_3 = hf(x_1 + h, y_1 + 2k_2 - k_1) = (0.2)f(0.4, 2.3254) = (0.2) \cdot (4.6509) = 0.9302\\
&y_2 = y_1 + \frac{1}{6}(k_1 + 4k_2 + k_3) = 1.4907 + \frac{1}{6}[0.5963 + 4(0.7155) + (0.9302)] = 2.2221\\
&y(0.4) = 2.2221
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_2, y_2) = (0.2)f(0.4, 2.2221) = (0.2) \cdot (4.4442) = 0.8888\\
&k_2 = hf(x_2 + \frac{h}{2}, y_2 + \frac{k_1}{2}) = (0.2)f(0.5, 2.6665) = (0.2) \cdot (5.333) = 1.0666\\
&k_3 = hf(x_2 + h, y_2 + 2k_2 - k_1) = (0.2)f(0.6, 3.4665) = (0.2) \cdot (6.9329) = 1.3866\\
&y_3 = y_2 + \frac{1}{6}(k_1 + 4k_2 + k_3) = 2.2221 + \frac{1}{6}[0.8888 + 4(1.0666) + (1.3866)] = 3.3124\\
&y(0.6) = 3.3124
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_3, y_3) = (0.2)f(0.6, 3.3124) = (0.2) \cdot (6.6248) = 1.325\\
&k_2 = hf(x_3 + \frac{h}{2}, y_3 + \frac{k_1}{2}) = (0.2)f(0.7, 3.9749) = (0.2) \cdot (7.9497) = 1.5899\\
&k_3 = hf(x_3 + h, y_3 + 2k_2 - k_1) = (0.2)f(0.8, 5.1673) = (0.2) \cdot (10.3347) = 2.0669\\
&y_4 = y_3 + \frac{1}{6}(k_1 + 4k_2 + k_3) = 3.3124 + \frac{1}{6}[1.325 + 4(1.5899) + (2.0669)] = 4.9377\\
&y(0.8) = 4.9377
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_4, y_4) = (0.2)f(0.8, 4.9377) = (0.2) \cdot (9.8753) = 1.9751\\
&k_2 = hf(x_4 + \frac{h}{2}, y_4 + \frac{k_1}{2}) = ((0.2)f(0.9, 5.9252) = (0.2)?(11.8504) = 2.3701\\
&k_3 = hf(x_4 + h, y_4 + 2k_2 - k_1) = (0.2)f(1, 7.7028) = (0.2)?(15.4055) = 3.0811\\
&y_5 = y_4 + \frac{1}{6}(k_1 + 4k_2 + k_3) = 4.9377 + \frac{1}{6}[1.9751 + 4(2.3701) + (3.0811)] = 7.3604\\
&y(1) = 7.3604
\end{aligned}$$

Essa aproximação é reazoável quando comparamos com a solução do problema em $y(1) = e^2 = 7.38906$.

7.2.4 Método RK-4

O método de quarta ordem Runge-Kutta (RK4), onde o ELT é $O(h^5)$, para aproximar a solução do problema de valor inicial $y'(x) = f(x, y); y(x_0) = y_0$, avalia o integrando, $f(x, y)$, quatro vezes por passo. Para o passo n,

$$\begin{aligned}
&k_1 = hf(x_{n-1}, y_{n-1}) \\
&k_2 = hf(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{k_1}{2}) \\
&k_3 = hf(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{k_2}{2}) \\
&k_4 = hf(x_{n-1} + h, y_{n-1} + k_3) \\
&y_n = y_{n-1} + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \qquad \text{(RK4)} \\
&\text{e} \\
&x_n = x_0 + n \cdot h
\end{aligned} \tag{7.13}$$

Exemplo 7.5

Considere o problema de valor inicial

$$y'(x) = 2y(x), y(0) = 0$$

cuja solução é $y(x) = e^{2x}$. O método RK4 será aplicado a este problema.

Suponha que queremos calcular o valor aproximado de $y(1)$ com $h = x_i - x_{i-1} = 0.2$. As aproximações para a solução nos pontos da malha (valores de x_i) usando o RK4 são:

$$\begin{aligned}
&k_1 = hf(x_0, y_0) = (0.2)f(0, 1) = (0.2) \cdot (2) = 0.4 \\
&k_2 = hf(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}) = (0.2)f(0.1, 1.2) = (0.2) \cdot (2.4) = 0.48 \\
&k_3 = hf(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}) = (0.2)f(0.1, 1.24) = (0.2)?(2.48) = 0.496 \\
&k_4 = hf(x_0 + h, y_0 + k_3) = (0.2)f(0.2, 1.496) = (0.2)?(2.992) = 0.5984 \\
&y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
&y_1 = 1 + \frac{1}{6}[0.4 + 2(0.48) + 2(0.496) + (0.5984)] = 1.4917 \\
&y(0.2) = 1.4917
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_1, y_1) = (0.2)f(0.2, 1.4907) = (0.2) \cdot (2.9813) = 0.5963 \\
&k_2 = hf(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}) = (0.2)f(0.3, 1.7901) = (0.2) \cdot (3.5802) = 0.716 \\
&k_3 = hf(x_1 + \frac{h}{2}, y_1 + \frac{k_2}{2}) = (0.2)f(0.3, 1.8497) = (0.2) \cdot (3.6995) = 0.7399 \\
&k_4 = hf(x_1 + h, y_1 + k_3) = 0.2)f(0.4, 2.2316) = (0.2) \cdot (4.4633) = 0.8927 \\
&y_2 = y_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
&y_2 = 1.4917 + \frac{1}{6}[0.5967 + 2(0.716) + 2(0.7399) + (0.8927)] = 2.2253 \\
&y(0.4) = 2.2253
\end{aligned}$$

$$\begin{aligned}
&k_1 = hf(x_2, y_2) = (0.2)f(0.4, 2.2253) = (0.2) \cdot (4.4505) = 0.8901 \\
&k_2 = hf(x_2 + \frac{h}{2}, y_2 + \frac{k_1}{2}) = (0.2)f(0.5, 2.6703) = (0.2) \cdot (5.3406) = 1.0681 \\
&k_3 = hf(x_2 + \frac{h}{2}, y_1 + \frac{k_2}{2}) = (0.2)f(0.5, 2.7593) = (0.2) \cdot (5.5187) = 1.1037 \\
&k_4 = hf(x_2 + h, y_2 + k_3) = (0.2)f(0.6, 3.329) = (0.2) \cdot (6.658) = 1.3316 \\
&y_3 = y_2 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
&y_3 = 2.2253 + \frac{1}{6}[0.8901 + 2(1.0681) + 2(1.1037) + (1.3316)] = 3.3195 \\
&y(0.6) = 3.3195
\end{aligned}$$

$$k_1 = hf(x_3, y_3) = (0.2)f(0.6, 3.3195) = (0.2) \cdot (6.639) = 1.3278$$
$$k_2 = hf(x_3 + \frac{h}{2}, y_3 + \frac{k_1}{2}) = (0.2)f(0.7, 3.9834) = (0.2) \cdot (7.9668) = 1.5934$$
$$k_3 = hf(x_3 + \frac{h}{2}, y_3 + \frac{k_2}{2}) = (0.2)f(0.7, 4.1162) = (0.2) \cdot (8.2324) = 1.6465$$
$$k_4 = hf(x_3 + h, y_3 + k_3) = (0.2)f(0.8, 4.966) = (0.2) \cdot (9.932) = 1.9864$$
$$y_4 = y_3 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
$$y_4 = 3.3195 + \frac{1}{6}[1.3278 + 2(1.5934) + 2(1.6465) + (1.9864)] = 4.9518$$
$$y(0.8) = 4.9518$$

$$k_1 = hf(x_4, y_4) = (0.2)f(0.8, 4.9518) = (0.2) \cdot (9.9036) = 1.9807$$
$$k_2 = hf(x_4 + \frac{h}{2}, y_4 + \frac{k_1}{2}) = (0.2)f(0.9, 5.9422) = (0.2) \cdot (11.8844) = 2.3769$$
$$k_3 = hf(x_4 + \frac{h}{2}, y_4 + \frac{k_2}{2}) = (0.2)f(0.9, 6.1403) = (0.2) \cdot (12.2805) = 2.4561$$
$$k_4 = hf(x_4 + h, y_4 + k_3) = (0.2)f(1, 7.4079) = (0.2) \cdot (14.8158) = 2.9632$$
$$y_5 = y_4 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
$$y_5 = 4.9518 + \frac{1}{6}[1.9807 + 2(2.3769) + 2(2.4561) + (2.9632)] = 7.3868$$
$$y(1) = 7.3868$$

Essa aproximação é uma boa quando comparamos com a solução do problema em $y(1) = e^2 = 7.38906$. O Método Runge-Kutta (RK4), produz um resultado melhor em menos etapas.

Capítulo Execícios

1. Considere o problema de valor inicial

$$y'(x) = 5 - x^4, \;\; y(0) = 3$$

cuja solução é $y(x) = 5x - \frac{x^5}{5} + 3$.

a) Encontre uma aproximação para $y(2)$ usando os métodos de Euler, RK-2, RK-3, e RK-4, com $h = 1$ e $h = 0.5$;

b) Compare seus resultados com a solução exata dada por $y(2) = 6.6$

2. Dada tabela abaixo:

Item	Equação diferencial	Condição Inicial	Encontrar $\tilde{y}$
a.	$y'(x) = -sen(x)$	$y(0) = 1$	y(1)
b.	$y'(x) = \frac{1}{x}$	$y(1) = 0$	y(3)
c.	$y'(x) = \frac{y}{x} \cdot ln(x)$	$y(1) = 2$	y(2)
d.	$y'(x) = 5x^4$	$y(-1) = -2$	y(1)
e.	$y'(x) = 3y + sen(x) - 2cos(x)$	$y(0) = 1$	y(1)

Resolva cada uma das EDO's pelos métodos de Euler e RK-4 com $h = 0.1$, $h = 0.01$, $h = 0.001$, e $h = 0.0001$ usando um software numérico.

Bibliografia

[1] C.W. Ueberhuber: Numerical Computation: Methods, software and analysis. Springer Berlin Heidelberg (1997) Vol. 1 474 pages

[2] J. Stoer, R. Bulirsh: Introduction to Numerical analysis. 2nd ed. Springer-Verlag, Berlin Heidelberg New York Tokio (1993)

[3] J.F. Traub, H. Wozniakowski: On the Optimal Solution of Large Linear Systems. J. Assoc. Comp. Mach. 31 (1984), pp. 545-559.

[4] R.W. Hamming, E.A. Feigenbaum Introduction to applied numerical analysis. McGraw-Hill, Inc New York (1971)

[5] R.E. Moore: Interval Analysis. Prentice Hall, Englewood Clifs, NJ, USA (1966)

[6] G. Alefeld, J. Herzberger: Introduction to Interval Computations. Academic Press, New (1966)

[7] U.W. Kulish and W.L. Miranker: The Arithmetic of the Digital Computers: A New Approach. SIAM Review **28**, 1 (1986)

[8] Rademacher, H, A.: On the accumulation of errors in processes of integration on higg-speed calculating machines. Procedings of a symposium on large-scale digital calculating machinary. Annals Comp. Labor. Havard Univ. **16** (1948) pp 176-185

[9] Sterbenz, P.H.: Floating Point Computation. Prentice Hall, Englewood Clifs, NJ, USA (1974)

[10] Bailey, D.H.: MPFUN - A portable High Performance Multiprecision Package. NASA Ames Tech. Report RUR-90-022, (1990)

[11] Goldeberg, D: What Every Computer Scientist Should Know About Floating-Point Arithimethic. ACM Computing Surveys, **23** (1991) pp 5-48

[12] C.W. Ueberhuber: Numerical Computation: Methods, software and analysis. Springer Berlin Heidelberg (1997) Vol. 2 495 pages

[13] Fröberg, C-E.: Introduction to numerical analysis. 2nd ed. Addison-Wesley Pub. Co. , Reading, Mass, , 1965

[14] Süli, E and Mayers, D. F.: An Introduction to Numerical Analysis. Cambridge University Press, 2003

[15] Conte, S.D. and Boor, C.: Elementary Numerical Analysis: An algorithmic approach. 3rd ed. McGraw-Hill Book Company, New York, 1980

CÁLCULO NUMÉRICO PARA ENGENHEIROS COM METODOLOGIAS NINJAS

conhecimentolivre.org/home
contato@conhecimentolivre.org
editoraconhecimentolivre

Sérgio Mário Lins Galdino
Jornandes Dias da Silva
Cícero José da Silva
Willames de Albuquerque Soares
Juan Carlos Oliveira de Medeiros

EDITORA CONHECIMENTO LIVRE