Cálculo Numérico em 24 aulas: uma abordagem para engenharia

Dr. Diego Eckhard

1º de Março de 2017

Conteúdo

Licença

Parte deste livro foi adaptado do livro **Cálculo Numérico - Um Livro Colaborativo**, que pode ser encontrado em:

- `https://www.ufrgs.br/numerico`
- `https://github.com/livroscolaborativos/CalculoNumerico`

e foi organizado por

- Dagoberto Adriano Rizzotto Justo - UFRGS
- Esequia Sauter - UFRGS
- Fabio Souto de Azevedo - UFRGS
- Leonardo Fernandes Guidi - UFRGS
- Matheus Correia dos Santos - UFRGS
- Pedro Henrique de Almeida Konzen - UFRGS

Prefácio

Este livro foi inicialmente escrito como notas de aula da disciplina Cálculo Numérico - MAT01169, ministrada na Universidade Federal do Rio Grande do Sul. Por ser um livro de notas de aula, cada capítulo é chamado de "aula" e normalmente é ministrado por mim em dois períodos (uma hora e quaranta minutos). O conteúdo é dividido em seis partes, com quatro aulas em cada parte, formando blocos de conteúdo. Historicamente estes blocos são ministrados na ordem encontrada neste livro, mas muitas modificações podem ser feitas. Normalmente a cada duas aulas de conteudo faço uma aula apenas com exercícios. As avaliações são feitas após duas partes, totalizando três avaliações.

Este livro não tem a pretensão de descrever os conteudos com profundidade. Sempre indico para meus alunos outras bibliografias para estudar esta matéria. Este livro contudo, tem se mostrado muito útil para os alunos revisarem a matéria, por apresentar os conteúdos de forma conscisa e na mesma ordem em que foram ministrados em aula. Espero que este livro seja útil para você!

Diego Eckhard

Parte I

Contas no computador

Aula 1

Sistemas de numeração

"E se o ser humano tivesse 8 dedos nas mãos?"

1.1 Conceitos

Definição 1.1 ***Número*** *é a idéia de quantidade que nos vem à mente quando contamos, ordenamos e medimos.*

Definição 1.2 ***Numeral*** *é toda representação de um número, seja falada ou escrita.*

Definição 1.3 ***Algarismo*** *é todo símbolo numérico que usamos para formar os numerais escritos.*

Definição 1.4 ***Sistema de numeração*** *é todo conjunto de regras para a produção sistemática de numerais. No caso de sistemas de numeração escrita, a produção de numerais é feita através de combinações de algarismos e eventuais símbolos não numéricos (como a vírgula).*

Tabela 1.1: Numerais e algarismos

Quantidade	Numeral em português	Sistema de numeração decimal	Algarismos arábicos
	zero	0	0
.	um	1	1
:	dois	2	2
:.	três	3	3
::	quatro	4	4
::.	cinco	5	5
:::	seis	6	6
:::.	sete	7	7
::::	oito	8	8
::::.	nove	9	9
:::::	dez	10	
::::::	doze	12	
:::::::	quatorze	14	
::::::::	dezesseis	16	

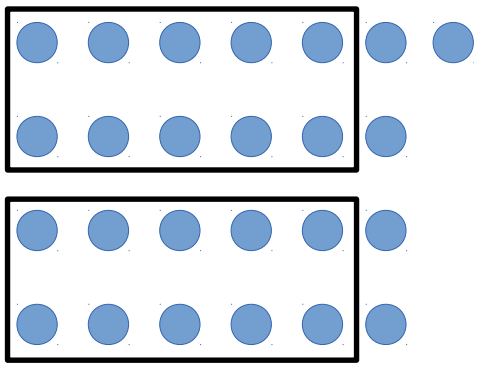


Figura 1.1: Sistema de numeração decimal

1.2 Sistema de numeração decimal

No sistema de numeração decimal as quantidades são agrupadas de dez em dez. Este sistema é o mais utilizado provavelmente porque o ser humano possui 10 dedos nas mãos.

Exemplo 1.1

$$25 = 2 \ \textit{dezenas} + 5 \ \textit{unidades}$$
$$25 = \mathbf{2} \cdot 10^1 + \mathbf{5} \cdot 10^0$$

Exemplo 1.2

$$293 = 2 \ \textit{centenas} + 9 \ \textit{dezenas} + 3 \ \textit{unidades}$$
$$293 = \mathbf{2} \cdot 10^2 + \mathbf{9} \cdot 10^1 + \mathbf{3} \cdot 10^0$$

1.3 Sistema de numeração octal

No sistema de numeração octal as quantidades são agrupadas de oito em oito. E se o ser humano tivesse 8 dedos nas mãos?

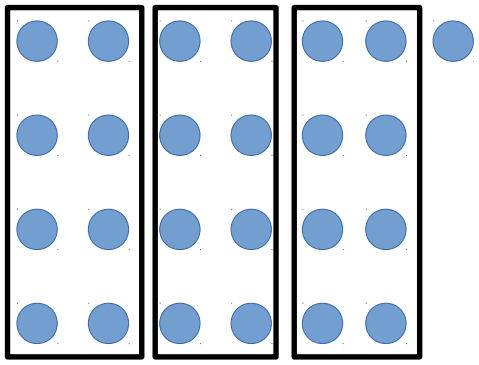


Figura 1.2: Sistema de numeração octal

Exemplo 1.3

$$25 = 3(\textit{grupos de } 8) + 1 \ \textit{unidade}$$
$$25 = \mathbf{3} \cdot 8^1 + \mathbf{1} \cdot 8^0 = (31)_8$$

Exemplo 1.4

$$293 = 4(\textit{grupos de } 64) + 4(\textit{grupos de } 8) + 5 \ \textit{unidade}$$
$$293 = \mathbf{4} \cdot 8^2 + \mathbf{4} \cdot 8^1 + \mathbf{5} \cdot 8^0 = (445)_8$$

1.4 Sistema de numeração binário

No sistema de numeração binário as quantidades são agrupadas de dois em dois. Os computadores utilizam apenas dois estados para armazenar informações de forma elétrica ou magnética.

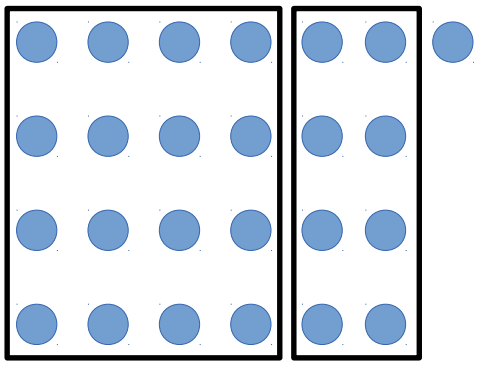


Figura 1.3: Sistema de numeração binário

Exemplo 1.5

$$\begin{aligned} 25 &= 1(\textit{grupo de } 2^4 \textit{ elementos}) + 1(\textit{grupo de } 2^3 \textit{ elementos}) + 0(\textit{grupos de } 2^2 \textit{ elementos }) \\ &\quad +0(\textit{grupos de } 2 \textit{ elementos}) + 1 \textit{ unidade} \\ 25 &= \mathbf{1} \cdot 2^4 + \mathbf{1} \cdot 2^3 + \mathbf{0} \cdot 2^2 + \mathbf{0} \cdot 2^1 + \mathbf{1} \cdot 2^0 = (11001)_2 \end{aligned}$$

Aula 2

Representação de números em um computador

"Existem 10 tipos de pessoas: as que entendem binário e as que não entendem."

2.1 Sistema de numeração de base b

Dado um número natural $b > 1$ e os algarismos 0,1,2,3,..., o número

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

representa o número decimal positivo

$$d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_1 \cdot b^1 + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots + d_{-m} \cdot b^{-m}$$

Exemplo 2.1

$$\begin{aligned}(325{,}33)_6 &= \mathbf{3} \cdot 6^2 + \mathbf{2} \cdot 6^1 + \mathbf{5} \cdot 6^0 + \mathbf{3} \cdot 6^{-1} + \mathbf{3} \cdot 6^{-2} \\ &= \mathbf{3} \cdot 36 + \mathbf{2} \cdot 6 + \mathbf{5} \cdot 1 + \mathbf{3} \cdot \frac{1}{6} + \mathbf{3} \cdot \frac{1}{36} \\ &= 108 + 12 + 5 + \frac{3}{6} + \frac{3}{36} = 125{,}5833333333 = 125{,}58\overline{3}\end{aligned}$$

2.2 Conversão de números de base binária para base decimal

Considere o número

$$(d_n\ d_{n-1}\ \cdots\ d_1\ d_0,\ d_{-1}\ d_{-2}\ \cdots\ d_{-m})_2.$$

A conversão para base decimal é feita por

$$d_n \cdot 2^n + d_{n-1} \cdot 2^{n-1} + \cdots + d_1 \cdot 2^1 + d_0 \cdot 2^0 + d_{-1} \cdot 2^{-1} + d_{-2} \cdot 2^{-2} + \cdots + d_{-m} \cdot 2^{-m}$$

Exemplo 2.2

$$\begin{aligned}(10{,}01)_2 &= \mathbf{1} \cdot 2^1 + \mathbf{0} \cdot 2^0 + \mathbf{0} \cdot 2^{-1} + \mathbf{1} \cdot 2^{-2} \\ &= \mathbf{1} \cdot 2 + \mathbf{0} \cdot 1 + \mathbf{0} \cdot 0{,}5 + \mathbf{1} \cdot 0{,}25 \\ &= 2 + 0{,}25 = 2{,}25\end{aligned}$$

2.3 Conversão de números de base decimal para base binária

Considere um número na base decimal D. Para realizar a conversão deste número para a base b é necessário encontrar os "dígitos" d_n tal que

$$D = \sum_{n=-\infty}^{\infty} d_n 2^n,$$

onde d_n é 0 ou 1.

Observe que a equação acima possui infinitas incógnitas d_n, mas permite apenas uma solução.

A conversão de números de base decimal para base binária é feita em duas partes. Primeiramente é convertida a parte inteira do número para depois ser feita a conversão da parte fracionária.

$$D = D_i + D_f.$$

Para cada parcela vamos apresentar um algoritmo.

2.3.1 Parcela Inteira

A conversão da parte inteira é feita realizando a divisão do número por 2 e armazenando o resto da divisão. Esta operação é repetida até que o resultado da divisão seja igual a 0. A representação binária é formada pelos restos tomados na ordem inversa a que foram obtidos.

Exemplo: $(401)_{10}$

dividendo	quociente	resto
401	200	$\underline{1}$
200	100	$\underline{0}$
100	50	$\underline{0}$
50	25	$\underline{0}$
25	12	$\underline{1}$
12	6	$\underline{0}$
6	3	$\underline{0}$
3	1	$\underline{1}$
1	0	$\underline{1}$

Portanto $(401)_{10} = (110010001)_2$.

2.3.2 Parcela Fracionária

Para converter a parte fracionaria, fazemos um processo de multiplicação por 2 e subtração da unidade quando o resultado for maior que um. O processo é repetido até que o resultado seja igual a 1,0. Os dígitos à esquerda do ponto decimal formam a representação binária do número, na ordem em que foram obtidos.

Exemplo: 0,640625

multiplicando	resultado
0,640625	$\underline{1}$,28125
0,28125	$\underline{0}$,5625
0,5625	$\underline{1}$,125
0,125	$\underline{0}$,25
0,25	$\underline{0}$,5
0,5	$\underline{1}$,0

Portanto $(0{,}640625)_{10} = (0{,}101001)_2$.

Juntando as incógnitas da parcela inteira com a parcela fracionária temos que $(401{,}640625)_{10} = (110010001{,}101001)_2$.

Os computadores armazenam os números e realizam as operações utilizando o sistema de numeração binário. Existem diversas maneiras de representar um número binário, onde as mais usuais são a representação por ponto-fixo e ponto-flutuante.

2.4 Representação de números reais em ponto-fixo

Os computadores representam os números utilizando uma certa quantidade de *bits* (binary digit). Atualmente os computadores são feitos, em sua maioria, com processadors de $64bits$, o que significa que são utilizados 64 dígitos em binário para representar um número real.

No sistema de ponto-fixo o ponto binário ocupa uma posição fixa (daí o nome) - existe uma quantidade pré-definida de dígitos à esquerda e a direita do ponto. O registrador do computador é dividido em três campos:

- s, sinal do número (|s| = 1 bit);
- e, dígitos à esquerda do ponto binário (|e| = 8 bits, por exemplo);
- d, dígitos à direita do ponto binário (|d| = 7 bits, por exemplo).

Por exemplo, o número $-11{,}75$ é representado em ponto-fixo como

1	00001011	1100000

Alguns números precisam ser arredondados, como por exemplo o número 1,2:

$$1{,}2 = (1{,}0011001100110011...) = (1, \overline{0011})$$

Arredondamento para baixo: Simplesmente descartam-se os dígitos em excesso.

$$(1{,}0011001100110011...)_2 \approx (1{,}0011001)_2$$

Arredondamento para mais próximo: Se o próximo dígito for 0 soma-se 0 no último dígito; se o próximo dígito for 1 soma-se 1 no último dígito.

$$(1{,}0011001|1)_2 \approx (1{,}0011001)_2 + (0{,}0000001)_2 = (1{,}0011010)_2$$

2.5 Números reais em ponto-flutuante - O padrão IEEE754

O padrão IEEE754 define regras para representação de números em ponto-flutuante. A representação foi criada com base na notação científica. Por exemplo:

$$1234{,}5 = 1{,}2345 \times 10^3$$

$$(1011{,}01)_2 = (1{,}01101)_2 \times 2^3$$

Portanto, um número real x é representado como

$$x = (-1)^s \times M \times 2^{C-BIAS}, \qquad 1 \leq M < 2$$

onde s representa o sinal, M é a mantissa, C é a característica e $BIAS$ é o deslocamento. Os números C e $BIAS$ são números inteiros enquanto que M é um número fracionário.

A característica é um número binário inteiro e é representada como

$$C = (c_m \cdots c_2 c_1 c_0)_2$$

A mantissa é um número binário entre *um* e *dois* e é representada como

$$M = (1, b_1\ b_2\ \cdots\ b_n)_2$$

O padrão descreve vários tipos de números, onde os mais usados são:

Tipo	n	m	BIAS	Total de bits
Meia precisão	10	4	15	1+10+5=16
Precisão simples	23	7	127	1+23+8=32
Precisão dupla	52	10	1023	1+52+11=64
Precisão quádrupla	112	14	16383	1+112+15=128

Um número em **precisão simples** pode ser armazenado no computador gravando os seguintes bits:

s	$c_7c_6c_5c_4c_3c_2c_1c_0$	$b_1b_2b_3\cdots b_{21}b_{22}b_{23}$

Por exemplo, $(-11{,}75)_{10} = (-1011{,}11)_2 = -(1{,}01111)_2 \times 2^3$.

Utilizando o sistema de ponto-flutuante precisão simples temos que

$$-(1{,}01111)_2 \times 2^3 = (-1)^1 \times (1{,}01111)_2 \times 2^{130-127},$$

portanto $s = 1$, a mantissa é 1,01111 e a característica é $130 = (10000010)_2$.

No computador este número é armazenado como

1	10000010	01111000000000000000000

2.5.1 Casos especiais

Alguns números especiais também podem ser gerados, e para isso é reservado $C = (000\cdots000)_2$ e $C = (111\cdots111)_2$ para representar estes números.

O número zero é um caso especial onde $s = 0$, $M = 0$ e $C = 0$ e o número é armazenado como

0	00000000	00000000000000000000000

Outro caso especial são $+\infty$ e $-\infty$ que são representados por $C = (111\cdots111)_2$ e $M = 0$.

2.6 Exercícios

Exercise 1 Converta para base decimal cada um dos seguintes números:

a) $(0100{,}111)_2$

b) $(0100{,}001)_2$

c) $(0000{,}001)_2$

d) $(1111{,}111)_2$

Exercise 2 Converta para base binária cada um dos seguintes números:

a) 52

b) 0,125

c) 854,5

d) 0,3

Exercise 3 Represente os números a seguir em **ponto-flutuante** de **meia-precisão** utilizando **arredondamento para o mais próximo**:

a) 0,567

b) 0,233

c) $-0{,}6785$

d) π

e) 99,76

Exercise 4 Considere a representação em ponto-fixo com $s = 1$, $e = 3$ e $d = 4$. Qual o maior número que pode ser representado? Qual o menor número? Qual o menor número positivo? Qual o maior número negativo?

Exercise 5 Considere a representação em ponto-flutuante meia precisão. Qual o maior número que pode ser representado? Qual o menor número? Qual o menor número positivo? Qual o maior número negativo?

Aula 3

Qualidade dos algoritmos

"Erros são relativos..."

3.1 Qualidade dos algoritmos

Nesta disciplina vamos criar algoritmos para resolver problemas numéricos, ou seja, criar algoritmos computacionais para resolver problemas de matemática. Ao criar um algoritmos precisamos avaliar dois pontos:

- Se o algoritmos gera a resposta certa;
- Quanto o algoritmo custa.

Nesta aula vamos analisar os erros dos algoritmos. Próxima aula vamos analisar o custo dos algoritmos.

3.2 Principais fontes de erros

- Dados de entrada: por exemplo, equipamentos de medição possuem precisão finita, acarretando erros nas medidas físicas.
- Erros de truncamento: ocorrem quando aproximamos um procedimento por um outro procedimento com menos etapas.

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \approx \sum_{n=0}^{M} \frac{x^n}{n!}$$

- Erros de arredondamento: são aqueles relacionados com as limitações que existem na forma de representar os números.

  Arredondamento para baixo - descarta os dígitos excedentes.

  π =3,14159265358979323846264338327950288419716939937510...
  π ≈3,141592

  Arredondamento para o mais próximo - procura pelo número mais próximo que contém a quantidade de dígitos desejada.

  π =3,14159265358979323846264338327950288419716939937510...
  π ≈3,141593

3.3 Medidas de Erros

Seja x um numero real e $\bar{x}$ sua aproximação, pode-se definir o erro absoluto e erro relativo.

3.3.1 Erro Absoluto

O erro absoluto é calculado como

$$|x - \bar{x}|$$

Observe que o erro absoluto apresenta a mesma unidade de medida da quantidade calculada.

Exemplo 3.1 *Considere $x = \frac{1}{3}kg$ e $\bar{x} = 0{,}333kg$. O erro absoluto é*

$$|x - \bar{x}| = \left|\frac{1}{3} - 0{,}333\right| = |0,\bar{3} - 0{,}333| = 0{,}000\bar{3} = 0,\bar{3} \cdot 10^{-3}kg$$

3.3.2 Erro Relativo

O erro relativo é calculado como

$$\frac{|x - \bar{x}|}{|x|}$$

Observe que o erro relativo é adimensional e pode ser escrito como porcentagem.

Exemplo 3.2 *Considere $x = \frac{1}{3}kg$ e $\bar{x} = 0{,}333kg$. O erro relativo é*

$$\frac{|x - \bar{x}|}{|x|} = \frac{\left|\frac{1}{3} - 0{,}333\right|}{\left|\frac{1}{3}\right|} = \frac{|0,\bar{3} - 0{,}333|}{|0,\bar{3}|} = \frac{0,\bar{3} \cdot 10^{-3}}{0,\bar{3}} = 10^{-3} = 0{,}1\%$$

3.4 Problemas causados pelos erros de arredondamento

3.4.1 Operações aritméticas de ponto-flutuante

Os números podem ser representados em **ponto-flutuante decimal** da seguinte forma:

$$123{,}45 = 0{,}12345 \times 10^3$$

Todos os dígitos estão no lado direito da vírgula. O número de dígitos é chamado de **dígitos significativos**.

Na soma e subtração, os expoentes dos dois operandos devem ser iguais. Para tal, ajustam-se as mantissas e os expoentes de tal forma a coincidir os expoentes utilizando sempre o maior deles. Uma vez feito o ajuste dos expoentes, basta calcular

$$(a \times r^p) \pm (b \times r^p) = (a \pm b) \times r^p$$

A multiplicação e a divisão são calculadas como

$$(a \times r^p) \times (b \times r^q) = (ab) \times r^{p+q}$$

$$(a \times r^p)/(b \times r^q) = (a/b) \times r^{p-q}$$

Portanto deve-se realizar as seguintes etapas:

- A operação é feita de forma "correta", i.e., o resultado é armazenado com o dobro do número de dígitos usados para armazenar cada operando;
- O resultado é normalizado;
- É feito o arredondamento, de forma que o resultado normalizado possa ser armazenado na palavra.

3.4.2 Cancelamento Catastrófico

Operações aritméticas entre números com representação finita podem fazer com que o resultado seja dominado pelos erros de arredondamento. Em geral, esse efeito, denominado cancelamento catastrófico, acontece quando fazemos a diferença entre números muito próximos entre si.

Exemplos

Exemplo 3.3 *Obtenha o resultado analítico da operação*

$$0{,}987624687925 - 0{,}987624 = 0{,}687926 \cdot 10^{-6}$$

e compare com o resultado obtido utilizando arredondamento com seis dígitos significativos.

Resultado analítico: $0{,}987624687925 - 0{,}987624 = 0{,}687926 \cdot 10^{-6}$

Arredondamento cada termo para baixo: $0{,}987624 - 0{,}987624 = 0$

Arredondamento cada termo para o mais próximo: $0{,}987625 - 0{,}987624 = 0{,}100000 \cdot 10^{-5}$

Observe que os resultados são bem distintos.

3.4.3 Instabilidade numérica

Ao desenvolver algoritmos numéricos é muito importante observar como os erros serão propagados, para prever a precisão do resultado.

Exemplo 3.4 *O número* $\frac{1}{3} = 0,\bar{3}$ *possui uma representação infinita tanto na base decimal quanto na base binária. Logo, quando representamos ele no computador é gerado um erro de arredondamento que denotaremos* ϵ*.*

Considere agora a seguinte sequência:

$$\begin{cases} x_0 = \frac{1}{3} \\ x_{n+1} = 4x_n - 1, \quad n \in \mathbb{N} \end{cases}$$

Observe que

$$\begin{aligned} x_0 &= \frac{1}{3} \\ x_1 &= 4\frac{1}{3} - 1 = \frac{1}{3} \\ x_2 &= 4\frac{1}{3} - 1 = \frac{1}{3} \\ x_3 &= 4\frac{1}{3} - 1 = \frac{1}{3}, \end{aligned}$$

ou seja, temos uma sequência constante igual a $\frac{1}{3}$*.*

Se calculamos no computador essa sequência, temos que incluir os erros de arredondamento, ou seja,

$$\begin{aligned} x_0 &= \frac{1}{3} + \epsilon \\ x_1 &= 4\left(\frac{1}{3} + \epsilon\right) - 1 = \frac{1}{3} + 4\epsilon \\ x_2 &= 4\left(\frac{1}{3} + 4\epsilon\right) - 1 = \frac{1}{3} + 4^2\epsilon \\ x_3 &= 4\left(\frac{1}{3} + 4^2\epsilon\right) - 1 = \frac{1}{3} + 4^3\epsilon \\ x_n &= \frac{1}{3} + 4^n\epsilon \end{aligned}$$

Portanto, o limite da sequência diverge!

3.5 Exercícios

Exercise 6 Arredonde os seguintes números (para o mais próximo) com 5 algarismos significativos:

- 56,781234
- 7812,563409

Exercise 7 Calcule os erros **absoluto** e **relativo** das aproximações $\bar{x}$ para x.

- $x = 3{,}1415926535$ e $\bar{x} = 3{,}141$
- $x = 1{,}00001$ e $\bar{x} = 1$

Exercise 8 Considere $x = 0{,}44523 \times 10^{-2}$ e $y = 0{,}22547 \times 10^{-3}$, em um sistema de ponto-flutuante decimal com cinco dígitos significativos. Qual o resultado de $x + y$, $x - y$, $x \times y$ e x/y?

Exercise 9 Arredonde os seguintes números com 4 algarismos significativos **truncando** e também **para o máis próximo**:

- 12345678
- 12,345678

Exercise 10 Considere $x = 2$ e $y = 30$.

a) Calcule $\bar{z} = \frac{x}{y}$ utilizando o sistema de ponto-flutuante decimal com cinco dígitos significativos.

b) Represente o resultado calculado $\bar{z}$, em número de máquina binário com meia precisão.

c) Calcule $z = \frac{x}{y}$ sem considerar aproximações.

d) Apresente o erro relativo e o erro absoluto entre z e o resultado calculado em ponto-flutuante $\bar{z}$.

Exercise 11

Considere $x = 1234$ e $y = 0.1234$.

a) Calcule $\bar{z} = x + y$ utilizando o sistema de ponto-flutuante decimal com três dígitos significativos.

b) Represente o resultado calculado $\bar{z}$, em número de máquina binário com meia precisão.

c) Calcule $z = x + y$ sem considerar aproximações.

d) Apresente o erro relativo e o erro absoluto entre z e o resultado calculado em ponto-flutuante $\bar{z}$.

Exercise 12 Considere o número $x = \sqrt{3}$ e a aproximação $\hat{x} = 1{,}73$. Sobre o erro absoluto e_a e o erro relativo e_r da aproximação, podemos afirmar que :

a) $e_a < 2 \times 10^{-3}$ e que $e_r > 0{,}0012$

b) $e_a = 0{,}0020$ e que $e_r = 0{,}0011$

c) $e_a < 2{,}1 \times 10^{-3}$ e que $e_r > 0{,}1\%$

d) $e_a > 2{,}1 \times 10^{-3}$ e que $e_r < 0{,}1\%$

e) Nenhuma das anteriores

Exercise 13 Considere o número $x = \sqrt{2}$ e a aproximação $\hat{x} = 1{,}41$. Sobre o erro absoluto e_a e o erro relativo e_r da aproximação, podemos afirmar que :

a) $e_a < 4{,}2 \times 10^{-3}$ e que $e_r > 0{,}029$

b) $e_a = 0{,}0042$ e que $e_r = 0{,}0030$

c) $e_a < 4{,}3 \times 10^{-3}$ e que $e_r > 2{,}9\%$

d) $e_a > 4{,}3 \times 10^{-3}$ e que $e_r < 2{,}9\%$

e) Nenhuma das anteriores

Aula 4

Erro de truncamento e Custo dos Algoritmos

"Time is money..."

4.1 Aproximações de funções

Considere uma função f real e infinitamente diferenciável. Está função pode ser escrita, utilizando a Série de Taylor como:

$$f(x+h) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)h^k}{k!}$$

Observe que a expressão acima não é *prática* pois envolve uma soma com infinitos termos. Muitas vezes esta soma é **truncada** gerando uma aproximação:

$$f(x+h) \approx \sum_{k=0}^{M} \frac{f^{(k)}(x)h^k}{k!}$$

Em geral, quanto maior o valor de M, melhor será a aproximação.

O Teorema de Taylor apresenta uma expressão para o erro desta aproximação. Segundo o teorema, existe um valor $t \in (x, x+h)$ tal que

$$f(x+h) = \sum_{k=0}^{M} \frac{f^{(k)}(x)h^k}{k!} + R(t)$$

onde $R(t)$ é o erro da aproximação dado por

$$R(t) = \frac{f^{M+1}(t)h^{M+1}}{(M+1)!}$$

Apesar de o valor t não ser conhecido, a expressão acima é bastante útil para estimar o tamanho do erro.

Vamos analisar um caso prático da função cosseno utilizando $x = 0$.

$$f(h) = cos(h) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)h^k}{k!} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!}h^{2k} = 1 - \frac{h^2}{2} + \frac{h^4}{24} + \cdots$$

Podemos truncar o somatório com $M = 4$:

$$f(h) \approx \sum_{k=0}^{4} \frac{f^{(k)}(0)h^k}{k!} = 1 - \frac{h^2}{2} + \frac{h^4}{24}$$

e portanto o erro será dado por

$$R(t) = cos^{(5)}(t)\frac{h^5}{5!} = -sen(t)\frac{h^5}{5!}$$

onde t é um número entre 0 e h. Apesar de não conhecermos o valor de t, sabemos que a função seno está sempre entre -1 e 1, logo podemos dizer queremos dizer que

$$|R(t)| < \left|\frac{h^5}{5!}\right| = \left|\frac{h^5}{120}\right|$$

A expressão acima apresenta um limite superior para o tamanho do erro. Observe que quanto maior o valor de h, maior será o erro. Podemos observar que o erro cresce quando h se distancia de zero na Figura 4.1.

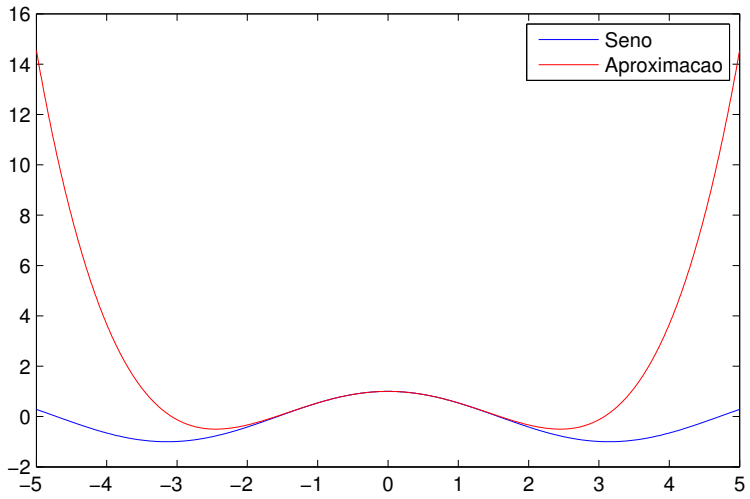


Figura 4.1: Função cosseno em azul e aproximação pela Série de Taylor em vermelho.

4.2 Notação Grande O - $\mathcal{O}$

Sejam f e g duas funções reais, podemos escrever:

$$f(x) = \mathcal{O}(g(x)) \text{ quando } x \to \infty$$

se e somente se existe uma constante positiva M e um número real x_0 tal que

$$|f(x)| \leq M|g(x)| \quad \forall x \geq x_0.$$

A notação também pode ser usada para mostrar o comportamento de f perto de um número real a:

$$f(x) = \mathcal{O}(g(x)) \text{ quando } x \to a$$

se e somente se existe constantes positivas M e δ e um número real a tal que

$$|f(x)| \leq M|g(x)| \quad \text{para } |x - a| \leq \delta.$$

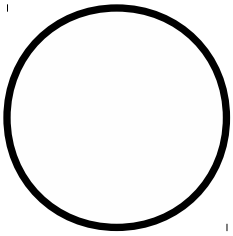
Esta notação é muito utilizada para descrever o tamanho do erro de funções. No caso da função cosseno, podemos dizer que o erro é dado por

$$R = \mathcal{O}(h^5) \text{ quando } h \to 0$$

O importante da expressão acima é descrever a relação do erro com o valor de h, que depende da quinta potência. Isto significa que se o valor de h for dividido por 10, o erro será diminuído para aproximadamente 1/100000 do erro anterior.

4.3 Custo de algoritmos

Calcule o perímetro da figura abaixo:



Teoricamente: $2\pi r$. Mas r vai ter erros e como calcular o π?

$$\pi = \frac{4}{1} - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \frac{4}{9} - \frac{4}{11} \cdots$$

O custo dos algoritmos pode ser medido por duas grandezas:

- **Tempo**: tempo que o algoritmo demora para rodar, normalmente medido em número de operações realizadas;
- **Memória**: quantidade de memória utilizada no computador.

Por exemplo, para calcular o perímetro acima poderíamos ter gravado no computador o número π com **muitas** casas, o que gastaria muita memória. Ou poderíamos calcular o número todas as vezes, o que levaria muito tempo.

4.4 Complexidade de algoritmos

A velocidade dos algoritmos usualmente é medida em número de operações matemáticas em ponto-flutuante realizadas por segundo (flops). Por exemplo:

$$5 + 4 \times 3$$

usa duas operações matemáticas. Quantas operações são usadas para calcular a multiplicação abaixo?

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \times \begin{pmatrix} 4 & 4 & 4 \\ 5 & 5 & 5 \\ 6 & 6 & 6 \end{pmatrix}$$

Três multiplicações por elemento mais duas somas: $(3 + 2)9 = 45$. Se fosse uma matriz $N \times N$? O total seria:

$$(N + (N - 1))(N \cdot N) = (2N - 1)N^2 = 2N^3 - N^2.$$

Exemplo 4.1 *Multiplicação de matrizes*

- *Para multiplicar uma matriz* 10×10*:* $2 \cdot (10^3) - 10^2 = 2000 - 100 = 1900$ *operações*
- *Para multiplicar uma matriz* 100×100*:* $2 \cdot (100^3) - 100^2 = 2000000 - 10000 = 1990000$ *operações*
- *Para multiplicar uma matriz* 1000×1000*:* $2 \cdot (1000^3) - 1000^2 = 2000000000 - 1000000 = 1999000000$ *operações*

Utilizando informação extra (linhas iguais, colunas iguais) podemos realizar apenas 3 multiplicações e 2 somas. Em uma matriz $N \times N$, o número de operações seria $N + (N - 1) = 2N - 1$.

Considere agora dois algoritmos que realizam a mesma operação em um vetor de tamanho N. Qual algoritmo é melhor?

- O número de operações do primeiro algoritmo é $f_1(N) = N^2 + 10N$.
- O número de operações do segundo algoritmo é $f_2(N) = 50N + 100$.

Neste exemplo o algoritmo 2 é melhor para valores grandes de N. Utilizando a notação de Grande O:

- $f_1(N) = \mathcal{O}(N^2)$ quando $x \to \infty$ - algoritmo de tempo quadrático
- $f_2(N) = \mathcal{O}(N)$ quando $x \to \infty$ - algoritmo de tempo linear

Os algoritmos podem ser classificados pelo sua velocidade utilizando a notação de Grande O:

Nome	Tempo de Execução
Tempo constante	$\mathcal{O}(1)$ quando $x \to \infty$
Tempo linear	$\mathcal{O}(N)$ quando $x \to \infty$
Tempo quadrático	$\mathcal{O}(N^2)$ quando $x \to \infty$
Tempo polinomial	$\mathcal{O}(N^k)$ quando $x \to \infty$
Tempo logaritmo	$\mathcal{O}(log(N))$ quando $x \to \infty$
Tempo fatorial	$\mathcal{O}(N!)$ quando $x \to \infty$

Logo, podemos dizer que a multiplicação de matrizes é um algoritmo de tempo polinomial.

Answer of exercise 1

a) 4,875

b) 4,125

c) 0,125

d) 15,875

Answer of exercise 2

a) $(110100)_2$

b) $(0{,}001)_2$

c) $(1101010110{,}1)_2$

d) $(0{,}0\overline{1001})_2$

Answer of exercise 3

a)

$$\begin{aligned} 0{,}567 &= (0.10010001001001101110100101111...)_2 \\ &= (1.0010001001001101110100101111...)_2 \times 2^{-1} \\ &\approx (1.0010001001)_2 \times 2^{-1} = (1.0010001001)_2 \times 2^{14-15} \\ &= (1.0010001001)_2 \times 2^{(1110)_2-15} \end{aligned}$$

$$0 \quad 01110 \quad 0010001001$$

b)

$$\begin{aligned} 0{,}233 &= (0.00111011101001011110001101010011...)_2 \\ &= (1.1101110100101111000110101001 1...)_2 \times 2^{-3} \\ &\approx (1.1101110101)_2 \times 2^{-3} = (1.1101110101)_2 \times 2^{12-15} \\ &= (1.1101110101)_2 \times 2^{(01100)_2-15} \end{aligned}$$

$$0 \quad 01100 \quad 1101110101$$

c)

$$\begin{aligned} -0{,}6785 &= -(0.10101101101100100010110100001 1...)_2 \\ &= -(1.0101101101100100010110100001 1...)_2 \times 2^{-1} \\ &\approx -(1.0101101110)_2 \times 2^{-1} = -(1.0101101110)_2 \times 2^{14-15} \\ &= -(1.0101101110)_2 \times 2^{(01110)_2-15} \end{aligned}$$

$$1 \quad 01110 \quad 0101101110$$

d)

$$\begin{aligned}\pi &= (11.00100100001111111011010101...)_2 \\ &= (1.10010010000111111011010101...)_2 \times 2^1 \\ &\approx (1.1001001000)_2 \times 2^1 = (1.1001001000)_2 \times 2^{16-15} \\ &= (1.1001001000)_2 \times 2^{(10000)_2-15}\end{aligned}$$

0	10000	1001001000

e)

$$\begin{aligned}99{,}76 &= (1100011.11000010100011111010111...)_2 = (1.10001111000010100011111010111...)_2 \times 2^6 \\ &\approx (1.1000111100)_2 \times 2^6 = (1.1000111100)_2 \times 2^{21-15} = (1.1000111100)_2 \times 2^{(10101)_2-15}\end{aligned}$$

0	10101	1000111100

Answer of exercise 4

Maior 7,9375

0	111	1111

Menor $-7{,}9375$

1	111	1111

Menor positivo 0,0625

0	000	0001

Maior negativo $-0{,}0625$

1	000	0001

Answer of exercise 5

Maior $(1{,}1111111111)_2 \times 2^{30-15} = (1{,}1111111111)_2 \times 2^{15} = 65504$

0	11110	1111111111

Menor $-(1{,}1111111111)_2 \times 2^{30-15} = (1{,}1111111111)_2 \times 2^{15} = -65504$

1	11110	1111111111

Menor positivo $(1{,}0000000000)_2 \times 2^{1-15} = (1{,}0000000000)_2 \times 2^{-14} = 2^{-14}$

0	00001	0000000000

Maior negativo $(1{,}0000000000)_2 \times 2^{1-15} = (1{,}0000000000)_2 \times 2^{-14} = -2^{-14}$

1	00001	0000000000

Answer of exercise 6

56,781 e 7812,6.

Answer of exercise 7

Answer of exercise 10

a)

$$x = 0.2 \times 10^1 \qquad y = 0.3 \times 10^2$$

$$\bar{z} = (0.2/0.3) \times 10^{1-2} = 0.6666666666 \times 10^{-1}$$

Arredondando

$$\bar{z} = 0.66667 \times 10^{-1} = 0.066667$$

b)

$$\bar{z} = 0.066667 = (0.0001000100010001...)_2$$

$$\bar{z} = 0.066667 = (1.000100010001...)_2 \times 2^{-4} = (1.000100010001...)_2 \times 2^{11-15}$$

$$11 = (1011)_2$$

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

c)

$$\bar{z} = 2/30 = 0.0666666\bar{6}$$

d)

$$Erro_{abs} = \|\bar{z} - z\| = \|0.066667 - 0.06\bar{6}\| = 3.33\bar{3} \times 10^{-7}$$

$$Erro_{rel} = \left\| \frac{\bar{z} - z}{z} \right\| = \left\| \frac{0.066667 - 0.06\bar{6}}{0.06\bar{6}} \right\| = 5 \times 10^{-6}$$

Answer of exercise 11

a)

$$x = 0.123400 \times 10^4 \quad y = 0.000012 \times 10^4$$

$$\bar{z} = 0.123412 \times 10^4$$

Arredondando

$$\bar{z} = 0.123 \times 10^4 = 1230$$

b)

$$1230 = (10011001110)_2 = (1.0011001110)_2 \times 2^10 = (1.0011001110)_2 \times 2^{25-15}$$

$$25 = (11001)_2$$

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

c)

$$z = x + y = 1234.1234$$

d)

$$Erro_{abs} = \|\bar{z} - z\| = \|1230 - 1234.1234\| = 4.1234$$

$$Erro_{rel} = \left\| \frac{\bar{z} - z}{z} \right\| = \left\| \frac{1230 - 1234.1234}{1234.1234} \right\| = 0.00334115697020245 = 3.34115697020245 \times 10^{-}3$$

Answer of exercise 12

Resposta c.

Answer of exercise 13

Nenhuma das anteriores.

Parte II

Raízes de funções

Aula 5

Método da bisseção

"Tinha uma raiz no meio do caminho..."

5.1 Equações não-lineares

Trata-se de determinar os valores de x para os quais é satisfeita a equação $f(x) = 0$. Esses valores são denominados **raízes** da função $f(x)$.

Os procedimentos usados na obtenção das raízes de uma função podem ser enquadrados dentro dos seguintes passos:

1. determinação do número de raízes da função;
2. isolamento das raízes em intervalos contendo uma única raiz;
3. cálculo, através de um processo numérico, do valor das raízes com a exatidão requerida.

5.2 Método gráfico: número de raízes e intervalos contendo uma única raiz

O gráfico da função $f(x)$ permite determinar as raízes da função $f(x)$, porém resulta mais prático colocar a equação na forma $g_1(x) = g_2(x)$ onde o desenho do gráfico das funções $g_1(x)$ e $g_2(x)$ pode ser realizado de forma mais simples que o desenho de $f(x)$. As abscissas dos pontos de interseção das curvas $g_1(x)$ e $g_2(x)$ são as raízes procuradas.

Exemplo 5.1 *Determinar o número de soluções da equação* $x + ln(x) = 2$ *e isolá-las em intervalos contendo uma única solução.*

Podemos escrever a equação na forma $ln(x) = 2 - x$ *e definir* $g_1(x) = ln(x)$ *e* $g_2(x) = 2 - x$. *Traçando os gráficos de* $g_1(x)$ *e* $g_2(x)$ *descobre-se que a equação possui apenas uma solução que está localizada no intervalo* (1,2).

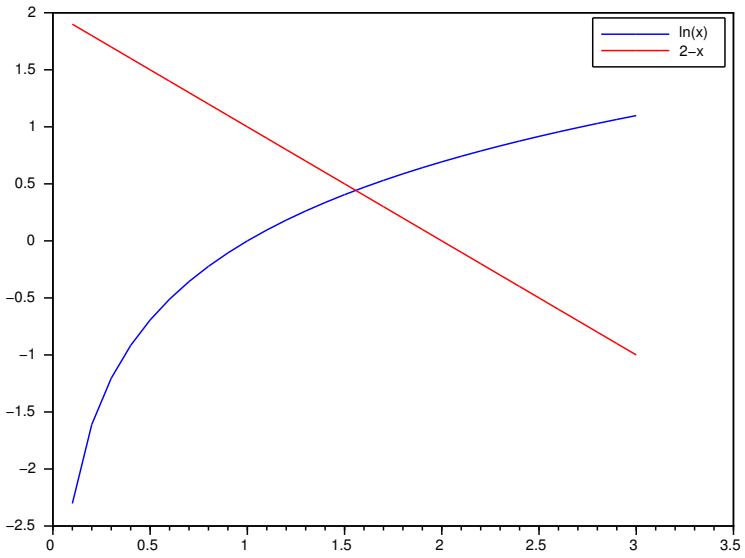


Figura 5.1: Método gráfico.

5.3 Método da Bisseção

Teorema 5.1 *Teorema de Bolzano*
Se $f(x)$ é contínua no intervalo $[a, b]$ e $f(a) \cdot f(b) < 0$, ou seja, a função $f(x)$ possui sinal diferente nos extremos do intervalo, então existe pelo menos uma raiz no intervalo $[a, b]$.

Exemplo 5.2 *Considere a função $f(x) = 2e^{-x} - 1$. No ponto $a = 0$ a função vale $f(0) = 1$ e em $b = 2$ a função vale $f(2) = -0.7293$. Como $f(0) \cdot f(2) < 0$ então existe uma raiz entre 0 e 2.*

O **Método da Bisseção** consistem em uma técnica para encontrar a raiz de uma função dentro de um intervalo $[a, b]$ tal que $f(a) \cdot f(b) \leq 0$. Para fazer isso, inicialmente o método procura pelo ponto central do intervalo:

$$p = \frac{a + b}{2}.$$

Se $f(p) = 0$ então a raiz foi encontrada, caso contrário avalia-se a função neste ponto para descobrir se a raiz da função está no intervalo $[a, p]$ ou no intervalo $[p, b]$. Para tanto, é feito o seguinte teste:

$$f(a) \cdot f(p) \leq 0.$$

Se $f(a) \cdot f(p) \leq 0$ então sabe-se que a raiz esta no intervalo $[a, p]$, caso contrário, a raiz está no intervalo $[p, b]$.

Note que ao fazer o teste acima, descobre-se um novo intervalo ($[a, p]$ ou $[p, b]$) que contém a raiz e que possui a metade do tamanho do intervalo inicial $[a, b]$. Pode-se repetir esta operação diversas vezes, e a cada iteração será encontrada uma aproximação melhor para a raiz da função.

Observe que se sabemos que existe uma raiz x_0 no intervalo $[a, b]$ e estimamos esta raiz por p, então o erro absoluto entre a raiz e a estimativa é dado por:

$$e = |x_0 - p| \leq \frac{b - a}{2}$$

Em cada iteração o erro é dividido por 2 e portanto temos uma estimativa para o tamanho do erro. Para melhor a estimativa (diminuir o erro) basta realizar mais iterações.

Tabela 5.1: Método da Bisseção

Iteração	a	b	p	$f(a)$	$f(b)$	$f(p)$	e
1	1	2	1,5	-	+	+	e<0,5
2	1	1,5	1,25	-	+	-	e<0,25
3	1,25	1,5	1,375	-	+		e<0,125
4							
5							

Exemplo 5.3 *Repita* 5 *vezes a operação acima para encontrar uma raiz da função* $f(x) = x^3 + 5x^2 - 12 = 0$ *no intervalo* [1,2].

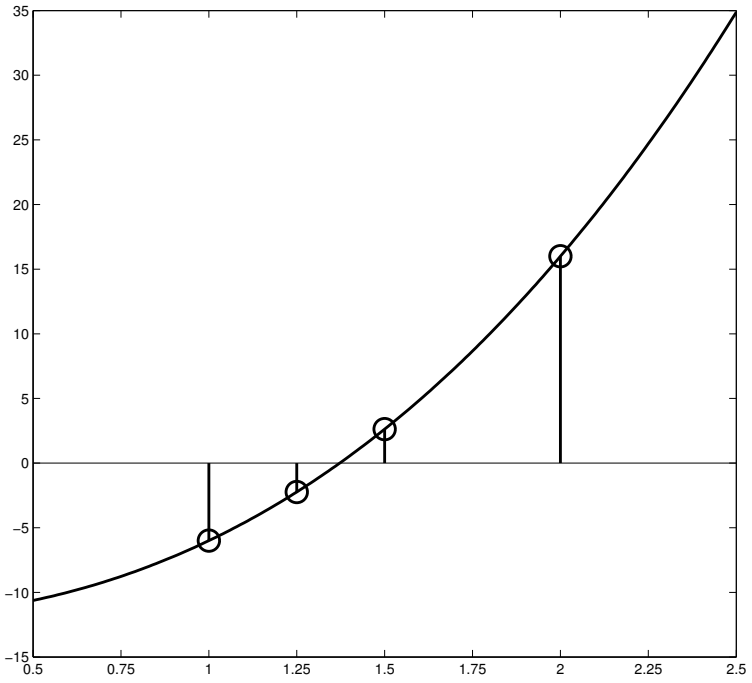


Figura 5.2: Método da bisseção.

5.4 Exercícios

Exercise 14 A equação $\frac{x}{15} = sen(x)$ possui quantas soluções?

Exercise 15 A equação $\frac{x}{15} - cos(x) = 0$ possui quantas soluções?

Exercise 16 Determinar o número de soluções e os intervalos contendo uma única solução utilizando o método gráfico. Utilize o método da bisseção para aproximar todas as soluções das equações abaixo.

- $e^{\frac{x}{2}} + x^2 - 3 = 0$
- $e^x + sen(x) = 2$

- $x^3 - x^2 + 3x - 2 = 0$
- $(x-3)^2 - 2cos(x-2) = 0$
- $\sqrt{x} - 2ln(x) = 1$

Exercise 17 Considere o método da bisseção para encontrar a raiz de uma função entre 103 e 104, sabendo que a função apresenta apenas uma raiz neste intervalo. O número mínimo de iterações N necessárias para obter uma aproximação com erro menor que 10^{-4} é

a) 13 iterações.

b) 14 iterações.

c) 15 iterações.

d) 16 iterações.

e) Nenhuma das anteriores

Exercise 18 Considere o método da bisseção para encontrar a raiz de uma função entre 13 e 14, sabendo que a função apresenta apenas uma raiz neste intervalo. O número mínimo de iterações N necessárias para obter uma aproximação com erro menor que 10^{-3} é

a) 9 iterações.

b) 10 iterações.

c) 11 iterações.

d) 12 iterações.

e) Nenhuma das anteriores

Exercise 19 Qual a solução da equação $cos(x) = x$ no intervalo entre 0 e 1 com 8 algarismos significativos?

Exercise 20 Qual a solução da equação $sen(x) = 2 - x$ no intervalo entre 1 e 2 com 7 algarismos significativos?

Aula 6

Método da Posição Falsa

"A raiz está próxima da corda."

6.1 Método da posição falsa

O **Método da posição falsa** é bastante parecido com o **Método da bisseção**.

A cada iteração o método procura por uma nova aproximação para a raiz da função. O método também divide o intervalo de busca $[a, b]$ em duas partes.

A principal diferença entre os métodos é que o **Método da bisseção** divide o intervalo $[a, b]$ exatamente ao meio, enquanto que o **Método da Posição Falsa**, divide o intervalo $[a, b]$ no ponto:

$$p = \frac{af(b) - bf(a)}{f(b) - f(a)}.$$

Este ponto p é o ponto em que a *corda* que passa entre os pontos $(a, f(a))$ e $(b, f(b)$ cruza o eixo da abcissa. Podemos ver este fato na Figura 6.1.

Exemplo 6.1 *Repita* 5 *iterações do Método da Posição Falsa para aproximar o valor de uma raiz da equação* $ln(x/2) = 0$ *no intervalo* [1,9].

Tabela 6.1: Método da posição falsa

Iteração	a	b	p	$f(a)$	$f(b)$	$f(p)$
1	1	9	3,52372	-	+	+
2	1	3,52372	2,38887	-	+	+
3	1	2,38887	2,1055	-	+	+
4	1	2,1055	2,02917	-	+	+
5	1	2,02917	2,00811	-	+	+

6.2 Critérios de Parada de Algoritmos

Os algoritmos de busca por raízes de funções procuram, em cada iteração, por uma aproximação x_n da solução x_* da equação $f(x) = 0$. Geralmente estes algoritmos não encontram o valor exato da solução, mas a cada iteração x_n se aproxima mais de x_*. Como os algoritmos podem nunca chegar na raiz é necessário utilizar um *critério de parada* para que os algoritmos não sejam executados indefinidamente.

Os três critérios de parada mais usuais são:

- Parada pelo número máximo de iterações;

  O algoritmo para após executar N iterações.

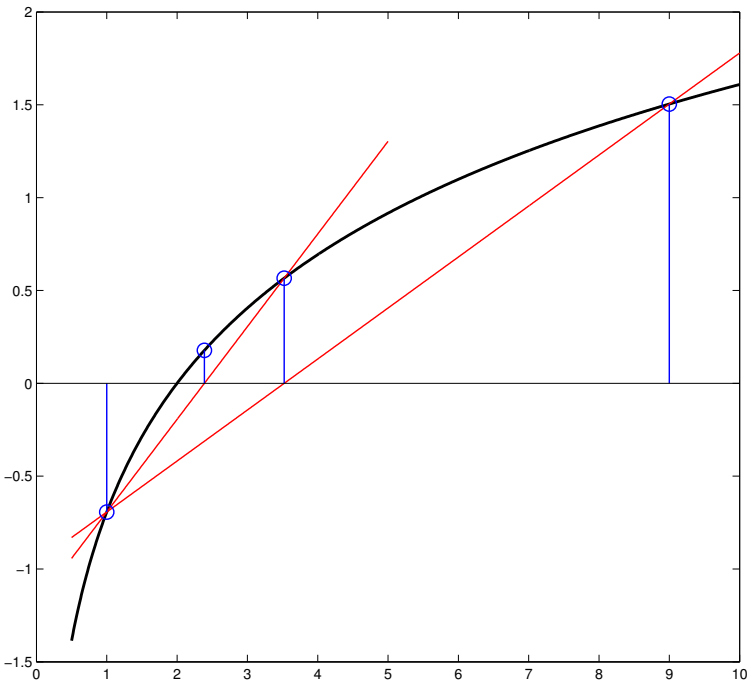


Figura 6.1: Exemplo 1

Exemplo: Algoritmo apresentado acima.

- Precisão do valor da função - $|f(x_n)| \leq e_1$;

  Observe que $f(x_*) = 0$ e portanto se x_n está próximo de x_* então $|f(x_n)|$ é bem pequeno.

  Exemplo:

```
if abs(f(p))<=e1 then break end
```

- Velocidade de convergência - $|x_{n+1} - x_n| \leq e_2$.

  Para a maioria dos algoritmos, quando x_n está próximo da raiz a velocidade do algoritmo é bastante baixa.

  Exemplo:

```
if abs(x-x0)<=e2 then break end
```

- Velocidade de convergência relativa - $|(x_{n+1} - x_n)/x_{n+1}| \leq e_3$.

  Para a maioria dos algoritmos, quando x_n está próximo da raiz a velocidade do algoritmo é bastante baixa.

  Exemplo:

```
if abs((x-x0)/x)<=e3 then break end
```

É possível também utilizar uma combinação dos critérios acima apresentados.

6.3 Exercícios

Exercise 21 A função $f(x) = x^2 + e^x - 2$ possui duas raízes. Qual o produto entre as duas raízes utilizando 4 algarismos significativos?

Exercise 22 Considere a equação $ln(x) + 5e^{-x/5} = \frac{\sqrt{x}}{10} + 3$. Calcule o produto de todas as soluções e arredonde para o número inteiro mais próximo.

Exercise 23 Qual o valor mínimo da função $f(x) = x^2 + e^x - 2$ utilizando 4 casas decimais?

Exercise 24 Qual o valor máximo da função $f(x) = ln(x) + 5e^{-x/5} - \frac{\sqrt{x}}{10} - 3$, utilizando 4 casas decimais?

Aula 7

Método da Secante e Método de Newton

"Rápido e rasteiro."

7.1 Método da Secante

O *Método da Secante* lembra bastante o método da Posição Falsa.

O método da Secante, utiliza o valor da função em dois pontos distintos para calcular uma nova aproximação para a raiz da função. A nova aproximação é calculada como o ponto em que a reta secante, que passa pelos dois pontos escolhidos, cruza o eixo das abscissas.

Por exemplo, considere a equação $f(x) = e^x - 2 = 0$. Suponha que os dois pontos escolhidos sejam $x_0 = 1{,}5$ e $x_1 = 2{,}5$, como pode ser visto na Figura 7.1.

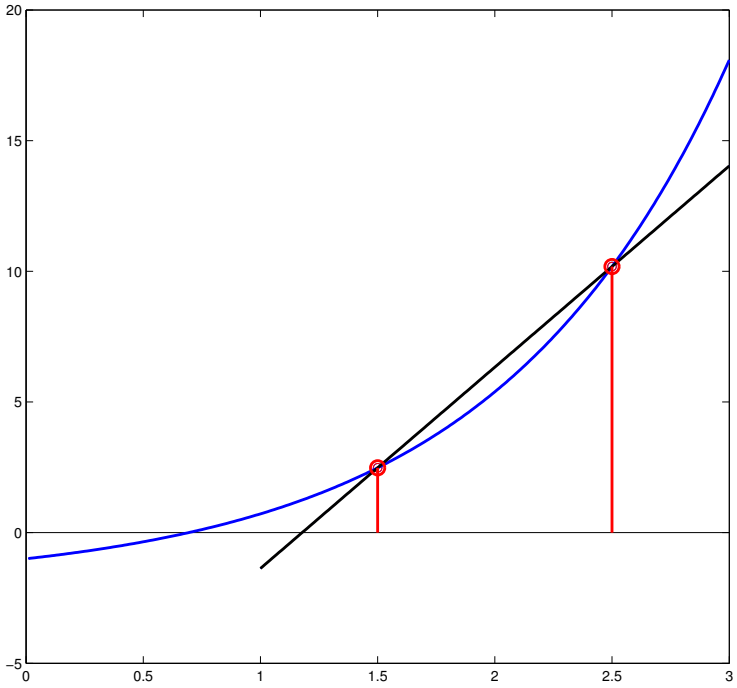


Figura 7.1: Método das secantes

A próxima aproximação será calculada como $x_2 = 1{,}1777$.

Para realizar o cálculo da nova aproximação, é necessário calcular a inclinação da reta secante, pela fórmula:

$$m = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

A nova aproximação será calculada por:

$$x_{n+1} = x_n - \frac{f(x_n)}{m}.$$

Contudo, esta fórmula apresenta problemas de *cancelamento catastrófico* porque ao chegar próxima da raiz, a atualização é bastante pequena. Uma formula que gera menos erros é dada por:

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{x_n f(x_{n-1}) - x_{n-1} f(x_n)}{f(x_n) - f(x_{n-1})}$$

Observação: Quando $f(x_n) = f(x_{n-1})$ ocorre uma divisão por zero, então devemos tomar cuidado para que isto não ocorra no algoritmo. Uma maneira de fazer isto é colocar um critério de parada no algoritmo quando esta condição ocorre.

Esta técnica pode ser repetida várias vezes, como no próximo exemplo.

Exemplo 7.1 *Encontre as raízes de* $f(x) = cos(x) - x$.

Pelo inspeção do gráfico sabemos que a função possui raiz entre $x = 0{,}7$ *e* $x = 0{,}8$.

Tabela 7.1: Método da Secante

n	x_{n-1}	x_n	m	x_{n+1}
1	0,7	0,8	$\frac{f(0,8)-f(0,7)}{0,8-0,7} = -1{,}6813548$	$0{,}8 - \frac{f(0,8)}{m} = 0{,}7385654$
2	0,8	0,7385654	$-1{,}6955107$	0,7390784
3	0,7385654	0,7390784	$-1{,}6734174$	0,7390851
4	0,7390784	0,7390851	$-1{,}6736095$	0,7390851

Observe que o algoritmo convergiu rapidamente.

7.2 Método de Newton

O *Método de Newton*, diferentemente dos outros métodos estudados, utiliza informação de apenas um ponto da função para calcular uma nova aproximação para a raiz da equação. A nova aproximação é calculada como o ponto em que a reta tangente cruza o eixo das abscissas.

Por exemplo, considere a função $f(x) = e^x - 2$. Suponha que o ponto escolhido seja $x_0 = 2$ como pode ser visto na Figura 7.2.

A próxima aproximação será calculada como $x_1 = 1{,}27067$

Para realizar o cálculo da nova aproximação, é necessário calcular a inclinação da reta tangente, pela fórmula:

$$m = f'(x_n)$$

A nova aproximação será calculada por:

$$x_{n+1} = x_n - \frac{f(x_n)}{m}.$$

Esta técnica pode ser repetida várias vezes, como no próximo exemplo.

Exemplo 7.2 *Encontre as raízes de* $f(x) = cos(x) - x$.

Pelo inspeção do gráfico sabemos que a função possui raiz próxima de $x = 0{,}8$.

Observe que o algoritmo convergiu rapidamente.

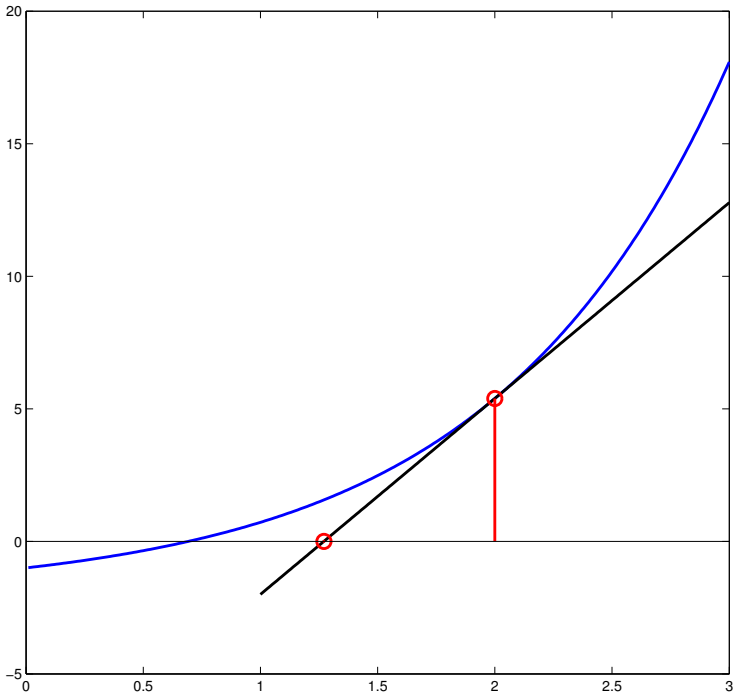


Figura 7.2: Método de Newton

Tabela 7.2: Método de Newton

Iteração	x_n	m	x_{n+1}
1	0,8	−1,7173560909	0,739853306370
2	0,739853306370	−1,67417957574	0,739085263405
3	0,739085263405	−1,673612125405	0,739085133215
4	0,739085133215	−1,673612029183	0,739085133215

7.3 Exercícios

Exercise 25 Utilize o Método da Secante e o Método de Newton para aproximar todas as soluções das equações abaixo.

1. $e^{\frac{x}{2}} + x^2 - 3 = 0$
2. $e^x + sen(x) = 2$
3. $x^3 - x^2 + 3x - 2 = 0$
4. $(x-3)^2 - 2cos(x-2) = 0$
5. $\sqrt{x} - 2ln(x) = 1$

Exercise 26 Encontre o ponto $(x; y)$ da curva $y = ln(x)$ mais próximo da origem.

Exercise 27 Encontre o ponto $(x; y)$ da curva $y = e^{-x}$ mais próximo da origem.

Aula 8

Convergência dos algoritmos

8.1 Métodos de Enquadramento

Observe que o Método da Bisseção na iteração i o erro e_i pode ser descrito como:

$$e_i = |x_0 - p_i| \leq \frac{b-a}{2^i}$$

Após muitas iterações o erro converge para zero com **ordem de convergência linear**.

$$\lim_{i \to \infty} \frac{|e_{i+1}|}{|e_i|^1} = 0{,}5$$

A ordem de convergência do método da Posição Falsa é $\frac{(1+\sqrt{5})}{2} \approx 1{,}618$, portanto geralmente o método converge mais rapidamente que o método da bisseção.

8.2 Métodos de Ponto-Fixo

A ordem de convergência do Método das Secantes é $\frac{1+\sqrt{5}}{2} \approx 1{,}618$, portanto geralmente o método converge mais rapidamente que o método da bisseção.

Já a ordem de convergência do Método de Newton é 2, portanto geralmente o método converge mais rapidamente que os outros vistos. Contudo, este método possui a desvantagem de necessitar do cálculo da derivada da função.

Answer of exercise 14

11.

Answer of exercise 15

9.

Answer of exercise 17

$i > \frac{log(1/e)}{log(2)}$. 14 iterações.

Answer of exercise 18

$i > \frac{log(1/e)}{log(2)}$. 10 iterações.

Answer of exercise 19

0,73908513

Answer of exercise 20

1,106060

Answer of exercise 21

$x1 = 0.537274449173857$ e $x2 = -1.315973777796290$. Logo, $x1 * x2 = -0.707039086592741$.

Answer of exercise 22

$x1 = 0.165956478975138$, $x2 = 6.354337972729073$, $x3 = 36.688186510535957$, $x4 = 2209.483047109994$, $x1 * x2 * x3 * x4 = 85483.33184521730$.

Answer of exercise 23

$-1{,}1728$

Answer of exercise 24

1,0059

Answer of exercise 25

1. 1,1183707 e $-1{,}5968611$
2. 0,4486719
3. 0,7152252
4. 1,6335428 e 3,4632767
5. 1 e 107,05321

Answer of exercise 26

$(0{,}652921; -0{,}42630)$

Answer of exercise 27

$(0{,}42630; 0{,}65292)$

Parte III

Matrizes

Aula 9

Sistemas de equações lineares

9.1 Sistemas de equações lineares

Trataremos de sistemas de equações algébricas lineares da seguinte forma:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m \end{aligned}$$

Observe que m é o número de equações e n é o número de incógnitas. Podemos escrever este problema na forma matricial

$$Ax = y$$

onde

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Daremos mais atenção ao caso $m = n$, isto é, quando a matriz A é quadrada.

9.2 Resolução de sistemas triangulares

São chamados de sistemas triangulares de equações lineares os sistemas em que a matriz A é triangular superior ou triangular inferior.

Podemos escrever este problema na forma matricial

$$Ax = y$$

onde

$$A = L = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \text{ ou } A = U = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

Um sistema triangular inferior pode ser resolvido simplesmente por

$$\begin{aligned} x_1 &= \frac{y_1}{a_{11}} \\ x_2 &= \frac{y_2 - a_{21}x_1}{a_{22}} \\ x_3 &= \frac{y_3 - a_{31}x_1 - a_{32}x_2}{a_{33}} \\ &\vdots \\ x_n &= \frac{y_n - \sum_{j=1}^{n-1} a_{nj}x_j}{a_{nn}} \end{aligned}$$

Um sistema triangular superior pode ser resolvido de maneira análoga.

$$\begin{aligned} x_n &= \frac{y_n}{a_{nn}} \\ x_{n-1} &= \frac{y_{n-1} - a_{(n-1)(n)}x_n}{a_{(n-1)(n-1)}} \\ &\vdots \\ x_1 &= \frac{y_1 - \sum_{j=2}^{n} a_{1j}x_j}{a_{11}} \end{aligned}$$

9.3 Eliminação gaussiana com pivotamento parcial

Lembramos que algumas operações feitas nas linhas de um sistema não alteram a solução:

1. Multiplicação de um linha por um número;

2. Troca de uma linha por ela mesma somada a um múltiplo de outra;

3. Troca de duas linhas.

O processo que transforma um sistema em outro com mesma solução, mas que apresenta uma forma triangular é chamado eliminação Gaussiana. A solução do sistema pode ser obtida fazendo substituição regressiva.

Exemplo 9.1 Eliminação Gaussiana com pivotamento parcial*: Resolva o sistema:*

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

Solução: Escrevemos a matriz completa do sistema:

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

Encontramos $1/2z = 1/2$*, ou seja,* $z = 1$*. Substituímos na segunda equação e temos* $y + 2z = 1$*, ou seja,* $y = -1$ *e, finalmente* $2x + y - z = 0$*, resultando em* $x = 1$*.*

Exemplo 9.2

$$\left[\begin{array}{ccc} 0 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{array}\right] \left[\begin{array}{c} x \\ y \\ z \end{array}\right] = \left[\begin{array}{c} 8 \\ 9 \\ 6 \end{array}\right]$$

Construímos a matriz completa:

$$\left[\begin{array}{ccc|c} 0 & 2 & 2 & 8 \\ 1 & 2 & 1 & 9 \\ 1 & 1 & 1 & 6 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & 2 & 2 & 8 \\ 1 & 1 & 1 & 6 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & 2 & 2 & 8 \\ 0 & -1 & 0 & -3 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & 2 & 2 & 8 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

Portanto $z = 1$*,* $y = \frac{8-2\cdot z}{2} = \frac{8-2\cdot 1}{2} = 3$*,*$x = \frac{9-2\cdot y-1\cdot z}{1} = \frac{9-2\cdot 3-1\cdot 1}{1} = 2$*.*

9.4 Exercícios

Exercise 28 Resolva o seguinte sistema de equações lineares

$$\begin{aligned} x + y + z &= 0 \\ x + 10z &= -48 \\ 10y + z &= 25 \end{aligned}$$

Usando eliminação gaussiana com pivoteamento parcial (não use o computador para resolver essa questão).

Exercise 29 Calcule a inversa da matriz

$$A = \left[\begin{array}{ccc} 1 & 2 & -1 \\ -1 & 2 & 0 \\ 2 & 1 & -1 \end{array}\right]$$

usando eliminação Gaussiana com pivotamento parcial.

Exercise 30 Considere o sistema:

$$\begin{aligned} 6x - 2y + 4z &= 14 \\ 5x - 3y + 5z &= 14 \\ 4x - 4y + 4z &= 8 \end{aligned}$$

Qual o valor de y que resolve o sistema?

a) 2

b) 1

c) 3

d) 4

e) Nenhuma das anteriores

Exercise 31 Considere o sistema:

$$\begin{aligned} 6x - 2y + 3z &= 5 \\ 5x - 3y + 5z &= 4 \\ 4x - 4y + 4z &= 0 \end{aligned}$$

Qual o valor de z que resolve o sistema?

a) 1

b) 2

c) 3

d) 4

e) Nenhuma das anteriores

Aula 10

Condicionamento de Sistemas Lineares

10.1 Motivação

Resolva o seguintes sistemas de equações lineares

$$\begin{cases} 71x + 41y = 100 \\ 52x + 30y = 70 \end{cases} \quad \text{e} \quad \begin{cases} 71x + 41y = 100 \\ 51x + 30y = 70 \end{cases}$$

A solução do primeiro problema é $x = -65$ e $y = 115$. A solução do segundo problema é $x = \frac{10}{3}$ e $y = -\frac{10}{3}$.

Note que os dois sistemas são bastante parecidos, onde apenas o fator que multiplica x na segunda equação é ligeiramente diferente, mas os dois sistemas possuem solução bastante distinta.

Diz-se que estes sistemas são **mal condicionados**, pois uma pequena variação nos parâmetros do sistema gera uma grande variação na solução. Estes sistemas são difíceis de resolver pois pequenos erros numéricos podem ocasionar grandes diferenças na solução obtida. Nesta aula vamos aprender como identificar estes sistemas, utilizando o **número de condicionamento**. Mas antes disso, vamos revisar conceitos sobre normas de vetores e matrizes.

10.2 Norma L_p de vetores

Definimos a norma L_p de vetores pertencentes ao $\mathbb{R}^n$, para $p \geq 1$ como

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots \ |x_n|^p)^{\frac{1}{p}}.$$

A norma L_∞ é definida como

$$\|x\|_\infty = \lim_{p \to \infty} \|x\|_p = \max_{1 \leq i \leq n} |x_i|$$

Propriedades: Seja $\lambda \in \mathbb{R}$ e $u, v \in \mathbb{R}^n$:

$$\begin{aligned} \|u\|_p &= 0 \Leftrightarrow u = 0 \\ \|\lambda u\|_p &= |\lambda| \cdot \|u\|_p \\ \|u + v\|_p &\leq \|u\|_p + \|v\|_p \quad \text{(desigualdade triangular)} \end{aligned}$$

Na Figura 10.1 são apresentados os conjuntos de pontos tais que $\|v\| = 1$. Em preto temos todos os pontos em que $\|v\|_1 = 1$. Em verde temos todos os pontos em que $\|v\|_2 = 1$. Em vermelho temos todos os pontos em que $\|v\|_3 = 1$. Em azul temos todos os pontos em que $\|v\|_\infty = 1$.

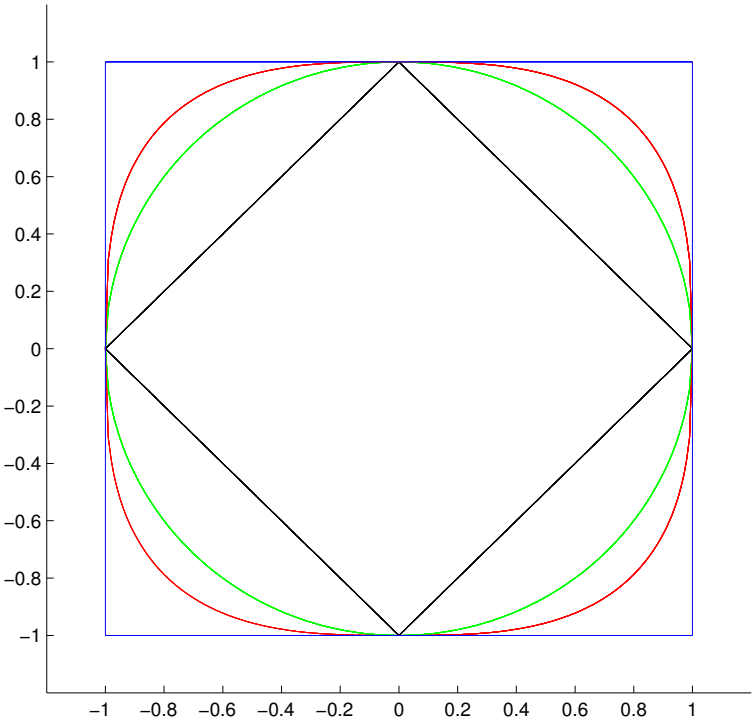


Figura 10.1: Normas de vetores

10.3 Norma L_p de matrizes

Definimos a norma operacional em L_p de uma matriz $A \in \mathbb{R}^{n\times n}$ como:

$$\|A\|_p = \max_{\|v\|_p=1} \|Av\|_p$$

ou seja, a norma p de uma matriz é o máximo valor assumido pela norma de Av entre todos os vetores de norma unitária.

Casos específicos:

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

$$\|A\|_2 = \sqrt{\lambda_{max}(AA^T)}$$

onde $\lambda_{max}(X)$ é o maior autovalor de X.

Propriedades: Seja $\lambda \in \mathbb{R}$, $v \in \mathbb{R}^n$, $A, B \in \mathbb{R}^{n\times n}$ e I a matriz identidade temos que

$$\begin{aligned}
\|A\|_p &= 0 \Leftrightarrow A = 0 \\
\|\lambda A\|_p &= |\lambda| \cdot \|A\|_p \\
\|A + B\|_p &\leq \|A\|_p + \|B\|_p \quad \text{(desigualdade triangular)} \\
\|Av\|_p &\leq \|A\|_p \cdot \|v\|_p \\
\|A \cdot B\|_p &\leq \|A\|_p \cdot \|B\|_p \\
\|I\|_p &= 1 \\
1 &= \|I\|_p = \|AA^{-1}\|_p \leq \|A\|_p \cdot \|A^{-1}\|_p \quad \text{se } A \text{ é inversível}
\end{aligned}$$

10.4 Número de condicionamento

Para analisar o condicionamento de um sistema descrito na forma

$$Ax = b$$

devemos avaliar o **número de condicionamento** da matriz A.

O número de condicionamento é definido como

$$k_p(A) = \|A\|_p \|A^{-1}\|_p.$$

Observe que o número de condicionamento pode ser definido para diferentes normas L_p, sendo que as mais utilizadas são as normas L_1, L_2 e L_∞.

Note que segundo as propriedades da norma L_p de matrizes, o número de condicionamento é sempre maior que 1.

Quanto mais próximo da unidade, melhor é o condicionamento do sistema de equações lineares, e por consequência, quanto maior o número de condicionamento, pior é o condicionamento do sistema.

Exemplo:

Calcule o número de condicionamento, para as normas L_1, L_2 e L_∞ para o sistema:

$$\begin{cases} 71x + 41y = 100 \\ 52x + 30y = 70 \end{cases}$$

Para este sistema, a matriz A é dada por

$$A = \begin{pmatrix} 71 & 41 \\ 52 & 30 \end{pmatrix}$$

$$k_1(A) = 6887{,}999999999792$$

$$k_2(A) = 5162{,}999806313514$$

$$k_\infty(A) = 6887{,}999999999792$$

Como o número de condicionamento é muito maior que a unidade é de se esperar que o sistema de equações seja mal condicionado.

10.5 Exercícios

Exercise 32 Considere o vetor $x = [1\ 2\ 3]$. Qual o valor de $||x||_3^3$?

a) 6

b) 36

c) 14

d) 0

e) Nenhuma das anteriores

Exercise 33 Considere o vetor $x = [1\ 2\ 3]$. Sobre a norma do vetor pode-se afirmar?

a) $3 \leq ||x||_5 \leq 3{,}25$

b) $3{,}25 \leq ||x||_5 \leq 3{,}5$

c) $3{,}5 \leq ||x||_5 \leq 3{,}75$

d) $3{,}75 \leq ||x||_5 \leq 4$

e) Nenhuma das anteriores

Exercise 34 Considere a matriz:

$$A = \begin{bmatrix} 5 & 6 \\ 6 & 7 \end{bmatrix}$$

Qual afirmativa esta correta?

a) $||A||_\infty < ||A||_1$

b) $||A||_1 \leq ||A^{-1}||_1$

c) $k_2(A) \neq k_2(A^{-1})$

d) $||A||_2 > ||A^{-1}||_2$

e) Nenhuma das anteriores

Aula 11

Métodos de Gauss-Jacobi e Gauss-Seidel

11.1 Método de Gauss-Jacobi

Um sistema de equações lineares pode ser escrito na forma

$$Ax = B. \tag{11.1}$$

Podemos escrever a matriz A na forma:

$$A = L + D + U$$

onde L é uma matriz estritamente triangular inferior (*lower*), D é uma matriz diagonal e U é uma matriz estritamente triangular superior.

O sistema de equações pode ser reescrito como

$$\begin{aligned} (L + D + U)x &= B & (11.2)\\ Dx &= -(L + U)x + B & (11.3)\\ x &= -D^{-1}(L + U)x + D^{-1}B & (11.4) \end{aligned}$$

O qual pode ser escrito como

$$x = Cx + E \tag{11.5}$$

onde $C = -D^{-1}(L + U)$ e $E = D^{-1}B$.

Observe que o sistema (11.1) e o sistema (11.5) possuem a mesma solução, que será denotada x_*.

O método de Gauss-Jacobi, consiste em aproximar a solução do sistema de equações iterativamente por

$$x^{k+1} = Cx^k + E \tag{11.6}$$

11.1.1 Exemplo

Considere o seguinte sistema de equações

$$\begin{cases} 10x_1 + 2x_2 - x_3 = 7 \\ 3x_1 + 20x_2 + 5x_3 = -12 \\ 1x_1 - 3x_2 + 10x_3 = 14 \end{cases}$$

Isolando x_1 na primeira equação, x_2 na segunda equação e x_3 na terceira equação chegamos em

$$\begin{cases} 10x_1 = -2x_2 + x_3 + 7 \\ 20x_2 = -3x_1 - 5x_3 - 12 \\ 10x_3 = -x_1 + 3x_2 + 14 \end{cases}$$

e

$$\begin{cases} x_1 = \frac{-2x_2+x_3+7}{10} \\ x_2 = \frac{-3x_1-5x_3-12}{20} \\ x_3 = \frac{-x_1+3x_2+14}{10} \end{cases}$$

Este sistema pode ser escrito na forma matricial como

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 & -2/10 & 1/10 \\ -3/20 & 0 & -5/20 \\ -1/10 & 3/10 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 7/10 \\ -12/20 \\ 14/10 \end{pmatrix}.$$

Vamos utilizar o Método de Gauss-Jacobi para estimar a solução do sistema de equações. Vamos iniciar o algoritmo com a estimativa

$$x^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Utilizando o método obtemos a nova aproximação como

$$x^1 = \begin{pmatrix} 0 & -2/10 & 1/10 \\ -3/20 & 0 & -5/20 \\ -1/10 & 3/10 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 7/10 \\ -12/20 \\ 14/10 \end{pmatrix} = \begin{pmatrix} 0{,}7 \\ -0{,}6 \\ 1{,}4 \end{pmatrix}.$$

Podemos repetir o método diversas vezes para melhorar a aproximação da solução

$$x^2 = \begin{pmatrix} 0{,}96 \\ -1{,}055 \\ 1{,}15 \end{pmatrix} \quad x^3 = \begin{pmatrix} 1{,}0260 \\ -1{,}0315 \\ 0{,}9875 \end{pmatrix} \quad x^4 = \begin{pmatrix} 1{,}0051 \\ -1{,}0008 \\ 0{,}9880 \end{pmatrix} \quad x^5 = \begin{pmatrix} 0{,}9990 \\ -0{,}9977 \\ 0{,}9993 \end{pmatrix}.$$

Note que o algoritmo está convergindo para a solução do problema

$$x_* = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

11.1.2 Convergência do método

Teorema 11.1 *O Método de Gauss-Jacobi converge para a solução do sistema de equações lineares se* $||C|| < 1$.

Prova 1 *Seja* x_* *a solução do sistema de equações, então*

$$x_* = Cx_* + E$$

A equação que rege o método é

$$x^{k+1} = Cx^k + E$$

Podemos então calcular o erro da aproximação

$$x_* - x^{k+1} = (Cx_* + E) - (Cx^k + E) = C(x_* - x^k)$$

e a norma do erro é

$$||x_* - x^{k+1}|| = ||C(x_* - x^k)||.$$

Utilizando propriedade referente à norma de um produto temos que

$$||x_* - x^{k+1}|| \leq ||C|| \cdot ||(x_* - x^k)||.$$

Conforme enunciado do teorema $||C|| < 1$ e portanto

$$||x_* - x^{k+1}|| < ||x_* - x^k||.$$

o que significa que a cada iteração a norma de erro diminui.

Logo,

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

Para que o Método de Gauss-Jacobi convirja para a solução do sistema de equações basta verificar se a norma de C é menor que um. Observe que não foi especificada qual norma deve ser usada, e portanto pode-se utilizar qualquer uma delas. Os testes mais utilizados são de $\|C\|_1$, $\|C\|_2$ e $\|C\|_\infty$.

11.1.3 Erro do método

Observe que

$$x_* - x^{k+1} = C(x_* - x^k)$$

Isolando x_*

$$x_* = x^{k+1} + C(x_* - x^k)$$

Subtraindo x^k nos dois lados

$$x_* - x^k = x^{k+1} - x^k + C(x_* - x^k)$$

Logo

$$\|x_* - x^k\| = \|x^{k+1} - x^k + C(x_* - x^k)\| \leq \|x^{k+1} - x^k\| + \|C(x_* - x^k)\| \leq \|x^{k+1} - x^k\| + \|C\|\|x_* - x^k\|$$

Isolando $\|x_* - x^k\|$:

$$\|x_* - x^k\| \leq \frac{1}{1 - \|C\|}\|x^{k+1} - x^k\|$$

Como $||x_* - x^{k+1}|| \leq ||C|| \cdot ||(x_* - x^k)||$,

$$\|x_* - x^{k+1}\| \leq \frac{\|C\|}{1 - \|C\|}\|x^{k+1} - x^k\|$$

11.2 Método de Gauss-Seidel

O método de Gauss-Jacobi calcula a cada iteração uma nova aproximação para a solução do problema por

$$x^{k+1} = Cx^k + E. \tag{11.7}$$

Cada linha de x^{k+1} é portanto calculado por

$$x_i^{k+1} = \sum_{j=1}^{n} C_{ij}x_j^k + E_i \qquad \text{para } i = 1,2,\ldots,n$$

onde i é o número da linha. Observe que são usados apenas valores **passados** da iteração x^k para calcular as novas linhas de x^{k+1}, mesmo quando **novos** valores já foram obtidos para algumas linhas de x^{k+1}.

Em contrapartida, o **Método de Gauss-Seidel** utiliza sempre os valores mais atuais para o calculo da nova aproximação, utilizando valores da aproximação que já foram obtidos. Desta maneira, espera-se que o método convirja mais rapidamente. Cada linha de x^{k+1} é calculada como

$$x_i^{k+1} = \sum_{j=1}^{i-1} C_{ij}x_j^{k+1} + \sum_{j=i}^{n} C_{ij}x_j^k + E_i \qquad \text{para } i = 1,2,\ldots,n$$

Matricialmente, o sistema de equações pode ser reescrito como

$$\begin{aligned} (L+D+U)x &= B && (11.8)\\ (D+L)x &= -Ux+B && (11.9)\\ x &= -(D+L)^{-1}Ux + D^{-1}B && (11.10)\end{aligned}$$

O qual pode ser escrito como

$$x = C_s x + E_s \tag{11.11}$$

onde $C_s = -D^{-1}(L+U)$ e $E_s = D^{-1}B$.

O método de Gauss-Seidel, consiste em aproximar a solução do sistema de equações iterativamente por

$$x^{k+1} = C_s x^k + E_s \tag{11.12}$$

11.2.1 Convergência do método

Teorema 11.2 *O Método de Gauss-Seidel converge para a solução do sistema de equações lineares se* $||C_s|| < 1$.

Teorema 11.3 *O Método de Gauss-Seidel converge também para a solução do sistema de equações lineares se A é simétrica definida positiva.*

11.3 Exercícios

Exercise 35 Considere um sistema de equações escrito no formato $x = Bx + C$ onde B é uma matriz com 3 linhas e 3 colunas, C é um vetor coluna com 3 elementos e x é um vetor coluna de incógnitas com 3 elementos.

a) Se $||B||_2 > 1$ o algoritmo de Gauss-Jacobi não consegue encontrar a solução do sistema.

b) Se $||B||_1 > 1$ o algoritmo de Gauss-Seidel não consegue encontrar a solução do sistema.

c) Este sistema de equações possui apenas uma solução.

d) Se I é a matriz identidade então $(I - B)x = C$.

e) Nenhuma das anteriores

Aula 12

Método da potência para cálculo de autovalores

12.1 Autovalores

Os autovalores e autovetores de uma matriz $A \in \mathbb{R}^{n\times n}$ podem ser calculados como solução da equação

$$Av = \lambda v$$

onde λ é um autovalor da matriz e v é o autovetor correspondente.

Podemos reescrever a equação como

$$(A - \lambda I)v = 0$$

que apenas possui solução para $v \neq 0$ se a matriz $(A - \lambda I)$ for singular. Portanto podemos procurar encontrar os autovalores de A como solução da equação

$$det(A - \lambda I) = 0.$$

A equação acima é polinomial de ordem n e pode ser resolvida utilizando diversas técnicas para solução de equações não-lineares. Resolvendo a equação podemos encontrar os n autovalores da matriz.

Exemplo 12.1 *Calcule os autovalores da matriz*

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -5 \end{bmatrix}.$$

$$A - \lambda I = \begin{bmatrix} 3-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & -5-\lambda \end{bmatrix}$$

$$det(A - \lambda I) = (3-\lambda)(2-\lambda)(-5-\lambda) = 0$$

$$\lambda_1 = 2$$
$$\lambda_2 = 3$$
$$\lambda_3 = -5$$

Fatos interessantes:

- *O maior autovalor é* $\lambda_2 = 3$
- *O menor autovalor é* $\lambda_3 = -5$
- *O maior autovalor em módulo é* $\lambda_3 = -5$
- *O menor autovalor em módulo é* $\lambda_1 = 2$

12.2 Método da potência

O método das potências é utilizado para calcular apenas o maior (em módulo) autovalor da matriz A.

12.2.1 Premissas

Vamos considerar que A é uma matriz inversível e que possui um autovalor dominante, de maneira que

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| > 0$$

Como A é inversível, o conjunto dos autovetores $\{v_j\}, j = 1,2,\ldots,n$ é linearmente independente e forma uma base para o espaço $\mathbb{R}^n$. Logo, qualquer $x \in \mathbb{R}^n$ pode ser escrito como

$$x = \sum_{j=1}^{n} \beta_j v_j.$$

12.2.2 O método

O método da potência consiste em utilizar o seguinte algoritmo para estimar o autovalor dominante de A:

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|}, \qquad \lambda_k = x_k^T A x_k.$$

Vamos mostrar que este algoritmo converge para o autovalor dominate de A para quase qualquer condição inicial x_0.

Note que segundo o algoritmo

$$x_1 = \frac{Ax_0}{\|Ax_0\|}, \quad x_2 = \frac{Ax_1}{\|Ax_1\|} = \frac{AAx_0}{\|AAx_0\|} = \frac{A^2x_0}{\|A^2x_0\|}, \quad x_3 = \frac{A^3x_0}{\|A^3x_0\|}, \quad x_k = \frac{A^kx_0}{\|A^kx_0\|}.$$

Observe também que $x_0 = \sum_{j=1}^{n} \beta_j v_j$ e portanto

$$A^k x_0 = A^k \sum_{j=1}^{n} \beta_j v_j = \sum_{j=1}^{n} \beta_j A^k v_j$$

Como os vetores v_j são autovetores de A:

$$\sum_{j=1}^{n} \beta_j A^k v_j = \sum_{j=1}^{n} \lambda_j^k \beta_j v_j = \lambda_1^k \left(\beta_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1} \right)^k \beta_j v_j \right)$$

Observe que $|\frac{v_j}{v_1}| < 1$ para $j > 1$ e portanto

$$\lim_{k\to\infty} \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1} \right)^k \beta_j v_j = 0$$

e para k grande

$$A^k x_0 \approx \lambda_1^k \beta_1 v_1 \quad e \quad \frac{A^k x_0}{\|A^k x_0\|} \approx \frac{\lambda_1^k \beta_1 v_1}{\|\lambda_1^k \beta_1 v_1\|}$$

Se $\beta_1 \neq 0$ então

$$\frac{\lambda_1^k \beta_1 v_1}{\|\lambda_1^k \beta_1 v_1\|} = \frac{v_1}{\|v_1\|} = v_1$$

e portanto o algoritmo converge para o autovetor associado ao autovalor dominante.

Agora, v_1 é um vetor unitário que também é um autovetor de A e portanto

$$v_1^T(Av_1) = v_1^T(\lambda_1 v_1) = \lambda_1 v_1^T v_1 = \lambda_1.$$

Portanto, para que o

12.3 Truques

12.3.1 Menor autovalor em módulo

Seja A é uma matriz não-singular; se λ é um autovalor de A, então λ^{-1} é autovalor de A^{-1}.

Exemplo 12.2 *Calcule os autovalores de $B = A^{-1}$ onde*

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -5 \end{bmatrix}.$$

$$\lambda_1 = \frac{1}{2}$$
$$\lambda_2 = \frac{1}{3}$$
$$\lambda_3 = -\frac{1}{5}$$

Utilize o método da potência para calcular o menor (em módulo) autovalor de A.

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -5 \end{bmatrix}.$$

Basta calcular o maior (em módulo) autovalor de A^{-1}. O menor autovalor de A será $1/\lambda$.

$$A^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & -1/5 \end{bmatrix}.$$

O maior autovalor em módulo de A^{-1} é $1/2$, portanto o menor autovalor em módulo de A é 2.

12.3.2 Maior e menor autovalor

Os autovalores de $(A + cI)$ onde c é um escalar são iguais aos autovalores de A acrescidos de c.

Exemplo 12.3 *Calcule os autovalores de $D = A + 7I$ onde*

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -5 \end{bmatrix}.$$

$$D = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

$$\lambda_1 = 10 = 3 + 7$$
$$\lambda_2 = 9 = 2 + 7$$
$$\lambda_3 = 2 = -5 + 7$$

Agora considere uma matriz A cujo maior autovalor em módulo seja λ_1 Sem perda de generalidade. Este autovalor será ou o maior autovalor de A ou o menor autovalor. Portanto o maior (ou menor) autovalor pode ser calculado utilizando o método da potência. Para calcular o outro (menor ou maior) autovalor, podemos calcular o maior autovalor em módulo de $A - \lambda_1$ e então somar λ.

Exemplo 12.4 *O maior autovalor em módulo de A é $\lambda_3 = -5$. Este autovalor é ou o menor ou o maior autovalor de A (de fato é o menor).*

Vamos calcular agora o maior autovalor em módulo de $A - \lambda_3$:

$$A + 5I = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

O maior autovalor em módulo é 8*. Somando este valor com λ_3 obtemos* 3*, e portanto o maior autovalor de A é* 3*.*

12.3.3 Desafio

Como utilizar o método da potência para calcular o segundo maior autovalor de $A = \begin{bmatrix} 7 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & -5 \end{bmatrix}$?

12.4 Exercícios

Exercise 36 Calcule o autovalor dominante de

- $\begin{pmatrix} 3 & 4 \\ 2 & -1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}$

Exercise 37 Calcule a norma L_2 de

- $\begin{pmatrix} 3 & 4 \\ 2 & -1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}$

Exercise 38 Considere a matriz:

$$A = \begin{bmatrix} 2{,}64 & -0{,}48 \\ -0{,}48 & 2{,}36 \end{bmatrix}$$

e que seus autovetores possuem norma $\mathcal{L}_2$ unitaria. É correto afirmar que:

a) A norma $\mathcal{L}_1$ do autovetor associado ao menor autovalor é 1,1.

b) A norma $\mathcal{L}_1$ do autovetor associado ao menor autovalor é 1.

c) A norma $\mathcal{L}_\infty$ do autovetor associado ao menor autovalor é 0,9.

d) A norma $\mathcal{L}_\infty$ do autovetor associado ao menor autovalor é 0,8.

e) Nenhuma das anteriores

Answer of exercise 28

$x = 2, y = 3, z = -5$

Answer of exercise 29

$$A^{-1} = \begin{bmatrix} -2 & 1 & 2 \\ -1 & 1 & 1 \\ -5 & 3 & 4 \end{bmatrix}$$

Answer of exercise 30

Resposta a) 2.

Answer of exercise 31

Resposta a) 1.

Answer of exercise 32

Resposta b) 36.

Answer of exercise 33

Resposta a) $3 \leq ||x||_5 \leq 3{,}25$. $||x||_5 = 3{,}0774$.

Answer of exercise 34

Resposta b). Observe que $||A||_1 = ||A^{-1}||_1 = 13$.

Answer of exercise 35

Resposta d).

Answer of exercise 36

4.4641 e 6.

Answer of exercise 37

5.0198 e 9.0125.

Answer of exercise 38

Resposta d).

Parte IV

Sistemas e Otimização

Aula 13

Sistemas de equações não-lineares

13.1 Sistemas de equações não-lineares

Nesta aula vamos aprender a resolver sistemas de equações não-lineares:

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \qquad\qquad \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{cases}$$

Considere por exemplo

$$\begin{cases} x_1^2 + x_2 - 3 = 0 \\ x_1 + x_2^2 - 5 = 0 \end{cases}$$

onde

$$f_1(x_1, x_2) = x_1^2 + x_2 - 3 \quad \text{e} \quad f_2(x_1, x_2) = x_1 + x_2^2 - 5.$$

Para resolver o sistema é necessário encontrar os valores de x_1 e x_2 que verificam o sistema de equações.

13.2 Método de Newton

O Método de Newton é um algoritmo iterativo utilizado para encontrar a solução de um sistema de equações não-lineares. O método se baseia em uma aproximação do sistema não-linear.

Considere uma função

$$f(x_1, x_2, \ldots, x_n)$$

fazendo a troca de variáveis

$$\begin{aligned} x_1 &= z_1 + h_1 \\ x_2 &= z_2 + h_2 \\ &\vdots \\ x_n &= z_n + h_n \end{aligned}$$

temos

$$f(x_1, x_2, \ldots, x_n) = f(z_1 + h_1, z_2 + h_2, \ldots, z_n + h_n)$$

A série de Taylor pode ser utilizada para obter uma aproximação de uma função:

$$f(z_1+h_1, z_2+h_2, \ldots, z_n+h_n) \approx f(z_1, z_2, \ldots, z_n) + h_1 \left.\frac{\partial f}{\partial x_1}\right|_{x_1=z_1} + h_2 \left.\frac{\partial f}{\partial z_2}\right|_{x_2=z_2} + \cdots + h_n \left.\frac{\partial f}{\partial z_n}\right|_{x_n=z_n}$$

Esta aproximação é boa quando os valores h são pequenos.

Utilizando a série de Taylor, podemos aproximar um sistema de equações não-lineares:

$$\left\{\begin{array}{l} f_1(z_1+h_1, z_2+h_2, \ldots, z_n+h_n) = 0 \\ f_2(z_1+h_1, z_2+h_2, \ldots, z_n+h_n) = 0 \\ \qquad\qquad \vdots \\ f_n(z_1+h_1, z_2+h_2, \ldots, z_n+h_n) = 0 \end{array}\right. \approx$$

$$\left\{\begin{array}{l} f_1(z_1, z_2, \ldots, z_n) + h_1 \left.\frac{\partial f_1}{\partial x_1}\right|_{x_1=z_1} + h_2 \left.\frac{\partial f_1}{\partial x_2}\right|_{x_2=z_2} + \cdots + h_n \left.\frac{\partial f_1}{\partial x_n}\right|_{x_n=z_n} = 0 \\ f_2(z_1, z_2, \ldots, z_n) + h_1 \left.\frac{\partial f_2}{\partial x_1}\right|_{x_1=z_1} + h_2 \left.\frac{\partial f_2}{\partial x_2}\right|_{x_2=z_2} + \cdots + h_n \left.\frac{\partial f_2}{\partial x_n}\right|_{x_n=z_n} = 0 \\ \qquad\qquad \vdots \\ f_n(z_1, z_2, \ldots, z_n) + h_1 \left.\frac{\partial f_n}{\partial x_1}\right|_{x_1=z_1} + h_2 \left.\frac{\partial f_n}{\partial x_2}\right|_{x_2=z_2} + \cdots + h_n \left.\frac{\partial f_n}{\partial x_n}\right|_{x_n=z_n} = 0 \end{array}\right.$$

Observe que o sistema aproximado pode ser escrito na forma matricial como:

$$\left[\begin{array}{cccc} \left.\frac{\partial f_1}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_1}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_1}{\partial x_n}\right|_{x_n=z_n} \\ \left.\frac{\partial f_2}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_2}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_2}{\partial x_n}\right|_{x_n=z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial f_1}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_1}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_1}{\partial x_n}\right|_{x_n=z_n} \end{array}\right] \left[\begin{array}{c} h_1 \\ h_2 \\ \vdots \\ h_n \end{array}\right] = \left[\begin{array}{c} -f_1(z_1, z_2, \ldots, z_n) \\ -f_2(z_1, z_2, \ldots, z_n) \\ \vdots \\ -f_n(z_1, z_2, \ldots, z_n) \end{array}\right]$$

Este sistema é linear e pode ser resolvido utilizando as técnicas conhecidas!

O Método de Newton consiste em "chutar" um valor para z_1, z_2, z_n e então calcular os valores dos hs utilizando o sistema acima. Uma nova (e melhor) aproximação para a solução do sistema será

$$\begin{array}{rcl} x_1 & = & z_1 + h_1 \\ x_2 & = & z_2 + h_2 \\ & \vdots & \\ x_n & = & z_n + h_n \end{array}$$

Esta nova solução pode ser usada como chute novamente, para refinar a solução, podendo ser repetido o procedimento diversas vezes.

O método pode ser descrito nos seguintes passos:

1. Chutar um valor inicial para z_1, z_2, z_n
2. Calcular a matriz

$$J(z) = \left[\begin{array}{cccc} \left.\frac{\partial f_1}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_1}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_1}{\partial x_n}\right|_{x_n=z_n} \\ \left.\frac{\partial f_2}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_2}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_2}{\partial x_n}\right|_{x_n=z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial f_1}{\partial x_1}\right|_{x_1=z_1} & \left.\frac{\partial f_1}{\partial x_2}\right|_{x_2=z_2} & \cdots & \left.\frac{\partial f_1}{\partial x_n}\right|_{x_n=z_n} \end{array}\right]$$

3. Calcular a vetor

$$F(z) = \left[\begin{array}{c} f_1(z_1, z_2, \ldots, z_n) \\ f_2(z_1, z_2, \ldots, z_n) \\ \vdots \\ f_n(z_1, z_2, \ldots, z_n) \end{array}\right]$$

4. Resolver o sistema
$$J(z)h = -F(z)$$
5. Calcular a solução $x = z + h$
6. Se a solução não for adequada, devemos chutar $z = x$ e ir para o passo 2.

Exercise 39 Resolva o sistema

$$\begin{cases} x_1^2 + x_2 = 3 \\ x_1 + x_2^2 = 5 \end{cases}$$

utilizando o Método de Newton.

Utilize como condição inicial
$$x = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Utilize como condição inicial
$$x = \begin{bmatrix} -5 \\ 5 \end{bmatrix}$$

Qual a diferença obtida?

Aula 14

Solução de problemas de otimização

14.1 Desafio

Encontre o mínimo global das seguintes funções:

1. $J_1(x, y) = x^4 + y^4$
2. $J_2(x, y) = e^x e^y (x^2 + y^2)$
3. $J_3(x, y) = sen(x) + sen(y) + \frac{x^2+y^2+xy}{100}$

14.2 Definições

Seja $J(\cdot) : \mathbb{R}^n \to \mathbb{R}$ uma função analítica em um domínio $\mathcal{D} \subseteq \mathbb{R}^n$. O gradiente desta função é um vetor coluna denotado

$$\nabla J(x) = \frac{\partial J(x)}{\partial x}.$$

A Hessiana desta função é uma matriz simétrica denotada

$$\nabla^2 J(x) = \frac{\partial^2 J(x)}{\partial x^2}.$$

Um ponto $x^1 \in \mathcal{D}$ é um *mínimo* de $J(\cdot)$ se $\exists \alpha > 0$ tal que $J(x) \geq J(x^1)$ para todo x que $\|x - x^1\| < \alpha$.

Um ponto $x^1 \in \mathcal{D}$ é um *mínimo isolado* se $\exists \alpha > 0$ tal que $J(x) > J(x^1)$ para todo x que $\|x - x^1\| < \alpha, x \neq x^1$.

Um ponto $x^1 \in \mathcal{D}$ é um *máximo* de $J(\cdot)$ se $\exists \alpha > 0$ tal que $J(x) \leq J(x^1)$ para todo x que $\|x - x^1\| < \alpha$.

Um ponto $x^1 \in \mathcal{D}$ é um *máximo isolado* se $\exists \alpha > 0$ tal que $J(x) < J(x^1)$ para todo x que $\|x - x^1\| < \alpha, x \neq x^1$.

Os máximos e mínimos são chamados de *extremos* da função.

Um ponto x^1 é chamado de *ponto crítico* de $J(\cdot)$ se $\nabla J(x^1) = 0$. Logo, qualquer extremo é um ponto crítico, mas nem todo ponto crítico é um extremo.

Se x^1 é um ponto crítico de $J(\cdot)$ e $\nabla^2 J(x^1) > 0$ então x^1 é um mínimo isolado de $J(\cdot)$.

Se x^1 é um ponto crítico de $J(\cdot)$ e $\nabla^2 J(x^1) < 0$ então x^1 é um máximo isolado de $J(\cdot)$.

14.3 Algoritmo do Gradiente

O algoritmo do gradiente é aquele descrito por

$$x_{k+1} = x_k - \gamma_k \nabla J(x_k),$$

ou seja, o algoritmo anda, a cada iteração na direção oposto ao gradiente. O tamanho do passo dado a cada iteração depende de γ_k.

Algumas regras utilizadas para a escolha de γ_k são:

- γ_k constante - $\gamma_k = c$, que gera convergência, embora seja lenta, se c for um valor pequeno;
- tamanho de passo constante - $\gamma_k = \frac{c}{\|\nabla J(x_k)\|}$, que gera convergência mais rápida para pontos distantes do mínimo, mas não garante convergência para o mínimo;

Note que o algoritmo "anda" na direção oposto ao gradiente, portanto se o tamanho do passo for pequeno, a cada passo o algoritmo encontra um novo ponto tal que o valor da função é menor que no passo anterior $J(x_{k+1}) < J(x_k)$.

14.3.1 Exemplo

Encontre o mínimo global de $J_1(x) = x_1^4 + x_2^4$. Esta função pode ser vista na Figura 14.1

O gradiente de $J_1(x)$ é

$$\nabla J_1(x) = \begin{pmatrix} \frac{\partial J_1}{\partial x_1} \\ \frac{\partial J_1}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 4x_1^3 \\ 4x_2^3 \end{pmatrix}$$

Vamos usar o algoritmo do gradiente com tamanho de passo $\gamma_k = 0.5/\|\nabla J(x_k)\|$, com condição inicial $x_0 = [-3\ 2]^T$.

O algoritmo convergiu rapidamente para uma região próxima do mínimo global, mas não convergiu exatamente para o mínimo.

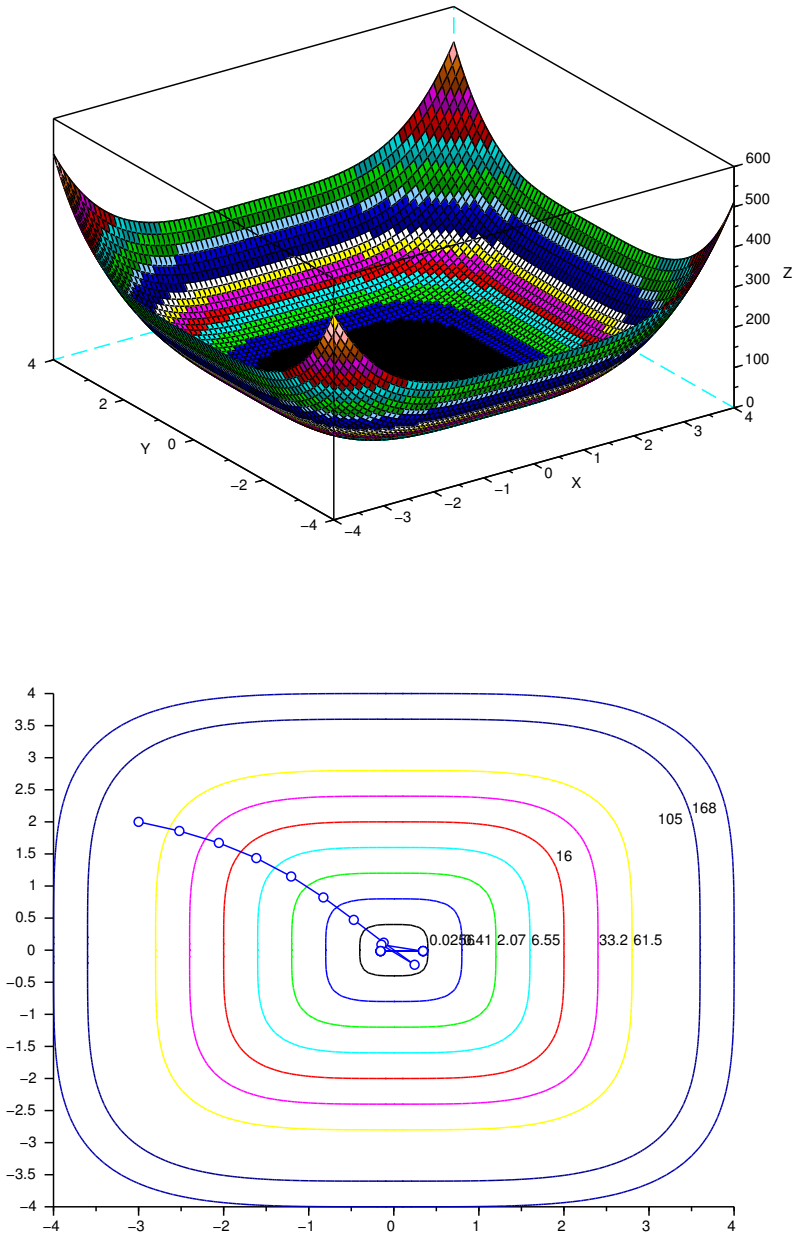


Figura 14.1: Função, curvas de nível da função e pontos do algoritmo

Aula 15

Ajuste de Curvas

15.1 Ajuste de Curvas

Ajuste de curvas consiste em encontrar os parâmetros de uma função de maneira que esta função represente da melhor maneira possível um conjunto de pontos.

Imagine que temos o seguinte conjunto de pontos:

x	1	2	3	4	5
y	1.1	1.9	3.2	4.9	5.1

Que pode ser representado pela Figura 15.1.

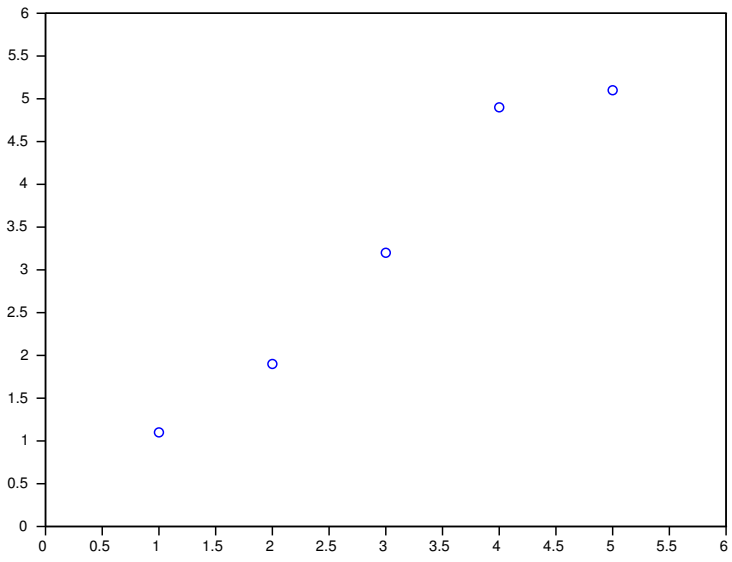


Figura 15.1: Pontos

Gostaríamos de encontrar a **reta** que melhor se encaixa nestes pontos. A equação da reta é

$$y = ax + b.$$

Para determinar os parâmetros a e b vamos resolver o seguinte problema de otimização:

$$\min_{a,b} J(a,b)$$

onde

$$J(a,b) = \sum_{i=1}^{5} \left(y_i - (ax_i + b)\right)^2.$$

Onde $J(a,b)$ é uma função que representa o erro entre a reta e os pontos. Quanto menor $J(a,b)$ melhor a reta descreve os pontos.

Para resolver este problema de otimização podemos utilizar várias técnicas. Uma delas consiste em igualar o gradiente de J a zero, e resolver o sistema de equações. Para o caso acima temos que

$$\nabla J(a,b) = \begin{pmatrix} \frac{\partial J}{\partial a} \\ \frac{\partial J}{\partial b} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{5}\left(y_i - (ax_i + b)\right)x_i \\ -2\sum_{i=1}^{5}\left(y_i - (ax_i + b)\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Observe que o conjunto de equações acima é linear e pode ser rescrito como

$$\begin{pmatrix} \sum_{i=1}^{5} x_i^2 & \sum_{i=1}^{5} x_i \\ \sum_{i=1}^{5} x_i & \sum_{i=1}^{5} 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{5} y_i x_i \\ \sum_{i=1}^{5} y_i \end{pmatrix}$$

Que no caso do exemplo é

$$\begin{pmatrix} 55 & 15 \\ 15 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 59.6 \\ 16.2 \end{pmatrix}$$

Cuja solução é

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1.1 \\ -0.06 \end{pmatrix}$$

15.1.1 Parábola

Caso seja desejado ajustar uma parábola aos pontos devemos definir a função como

$$y = ax^2 + bx + c.$$

E a função objetivo como

$$J(a,b,c) = \sum_{i=1}^{n} \left(y_i - (ax_i^2 + bx_i + c)\right)^2.$$

Para resolver o problema de otimização podemos novamente igualar o gradiente a zero:

$$\nabla J(a,b,c) = \begin{pmatrix} \frac{\partial J}{\partial a} \\ \frac{\partial J}{\partial b} \\ \frac{\partial J}{\partial c} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n}\left(y_i - (ax_i^2 + bx_i + c)\right)x_i^2 \\ -2\sum_{i=1}^{n}\left(y_i - (ax_i^2 + bx_i + c)\right)x_i \\ -2\sum_{i=1}^{n}\left(y_i - (ax_i^2 + bx_i + c)\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Novamente obtivemos um sistema de equações lineares que pode se escrito como

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i x_i^2 \\ \sum_{i=1}^{n} y_i x_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

15.1.2 Caso geral linear

Sempre que a função da curva for afim nas parâmetros, o problema de otimização pode ser resolvido utilizando um sistema de equações lineares. Nestes casos a função deve ser escrita como

$$y = k_1 f_1(x) + k_2 f_2(x) + \ldots + k_m f_m(x) = \sum_{i=j}^{m} k_j f_j(x)$$

Neste caso a função objetivo será

$$J(k_1, k_2, \ldots, k_m) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} k_j f_j(x) \right)^2.$$

Para resolver o problema de otimização podemos novamente igualar o gradiente a zero:

$$\nabla J = \begin{pmatrix} \frac{\partial J}{\partial k_1} \\ \frac{\partial J}{\partial k_2} \\ \vdots \\ \frac{\partial J}{\partial k_m} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} k_j f_j(x)\right) f_1(x) \\ -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} k_j f_j(x)\right) f_2(x) \\ \vdots \\ -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} k_j f_j(x)\right) f_m(x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Novamente obtivemos um sistema de equações lineares que pode se escrito como

$$\begin{pmatrix} \sum_{i=1}^{n} f_1(x)f_1(x) & \sum_{i=1}^{n} f_1(x)f_2(x) & \dots & \sum_{i=1}^{n} f_1(x)f_m(x) \\ \sum_{i=1}^{n} f_2(x)f_1(x) & \sum_{i=1}^{n} f_2(x)f_2(x) & \dots & \sum_{i=1}^{n} f_2(x)f_m(x) \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} f_m(x)f_1(x) & \sum_{i=1}^{n} f_m(x)f_2(x) & \dots & \sum_{i=1}^{n} f_m(x)f_m(x) \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i f_1(x) \\ \sum_{i=1}^{n} y_i f_2(x) \\ \vdots \\ \sum_{i=1}^{n} y_i f_m(x) \end{pmatrix}$$

15.1.3 Caso geral não-linear

No caso em que a função não é afim nos parâmetros que desejamos encontrar, o problema de otimização se torna mais complicado pois precisamos resolver um conjunto de equações não-lineares. Para tanto, podemos utilizar o **algoritmo do gradiente** e o **algoritmo de Newton**.

Em alguns casos específicos podemos realizar uma mudança de variável para "linearizar o problema". Vamos ver alguns casos:

- $y = ab^x$

  Calculando o logaritmo nos dois lados da equação temos que

  $$ln(y) = ln(ab^x) = ln(a) + ln(b^x) = ln(a) + xln(b)$$

  Fazendo as mudanças de variáveis $z = ln(y)$, $k_1 = ln(a)$ e $k_2 = ln(b)$ obtemos

  $$z = k_1 + k_2 x$$

  que é afim nas variáveis k_1 e k_2.

- $y = ax^b$

  Calculando o logaritmo nos dois lados da equação temos que

  $$ln(y) = ln(ax^b) = ln(a) + ln(x^b) = ln(a) + bln(x)$$

  Fazendo as mudanças de variáveis $z = ln(y)$, $k_1 = ln(a)$ e $k_2 = b$ obtemos

  $$z = k_1 + k_2 ln(x)$$

  que é afim nas variáveis k_1 e k_2.

- $y = \frac{1}{ax+b}$

  Fazendo as mudanças de variáveis $z = \frac{1}{y}$, $k_1 = a$ e $k_2 = b$ obtemos

  $$z = k_1 x + k_2$$

  que é afim nas variáveis k_1 e k_2.

- $y = \frac{1}{ax^2+bx+c}$

  Fazendo as mudanças de variáveis $z = \frac{1}{y}$, $k_1 = a$, $k_2 = b$, $k_3 = c$ obtemos

  $$z = k_1 x^2 + k_2 x + k_3$$

  que é afim nas variáveis k_1,k_2 e k_3.

- $y = ce^{ax^2+bx}$

  Calculando o logaritmo nos dois lados da equação temos que

  $$ln(y) = ln(c) + ax^2 + bx$$

  Fazendo as mudanças de variáveis $z = ln(y)$, $k_1 = a$, $k_2 = b$, $k_3 = ln(c)$ obtemos

  $$z = k_1x^2 + k_2x + k_3$$

  que é afim nas variáveis k_1,k_2 e k_3.

- $y = \frac{x}{ax+b}$

  Fazendo as mudanças de variáveis $z = \frac{x}{y}$, $k_1 = a$ e $k_2 = b$ obtemos

  $$z = k_1x + k_2$$

  que é afim nas variáveis k_1 e k_2.

15.2 Exemplo

Considere o seguinte conjunto de pontos:

x	1	2	3	4	5
y	1.666666666667	0.909090909091	0.555555555556	0.370370370370	0.263157894737

Que pode ser visto na Figura 15.2.

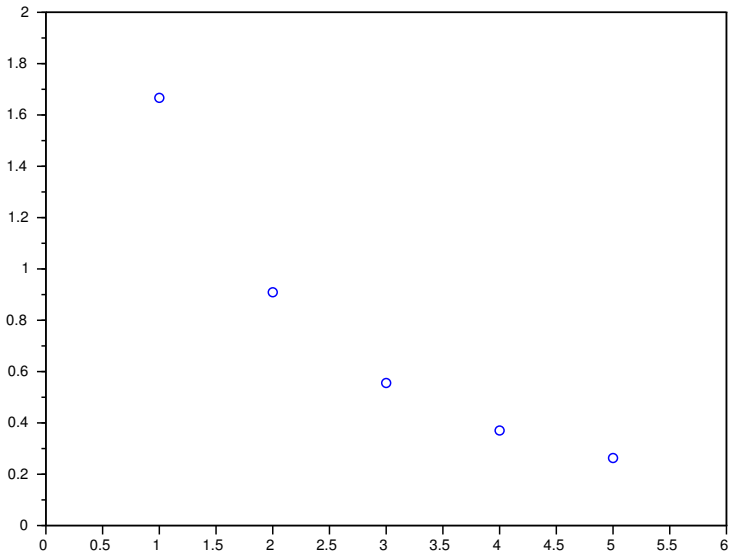


Figura 15.2: Exemplo

Gostaríamos de encontrar os parâmetros da função

$$y = \frac{1}{ax^2 + bx + c}$$

que melhor descrevem estes pontos.

Fazendo a mudança de variável $z = \frac{1}{y}$, obtemos

$$z = ax^2 + bx + c$$

E então o conjunto de pontos utilizados será

x	1	2	3	4	5
z	0.6	1.1	1.8	2.7	3.8

Para encontrar os parâmetros a, b e c temos que resolver

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} z_i x_i^2 \\ \sum_{i=1}^{n} z_i x_i \\ \sum_{i=1}^{n} z_i \end{pmatrix}$$

$$\begin{pmatrix} 1958. & 450. & 110. \\ 450. & 110. & 30. \\ 110. & 30. & 10. \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 318.8 \\ 76. \\ 20. \end{pmatrix}$$

cuja solução é

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \end{pmatrix}$$

portanto

$$y = \frac{1}{0.1x^2 + 0.2x + 0.3}$$

que pode ser vista na Figura 15.3.

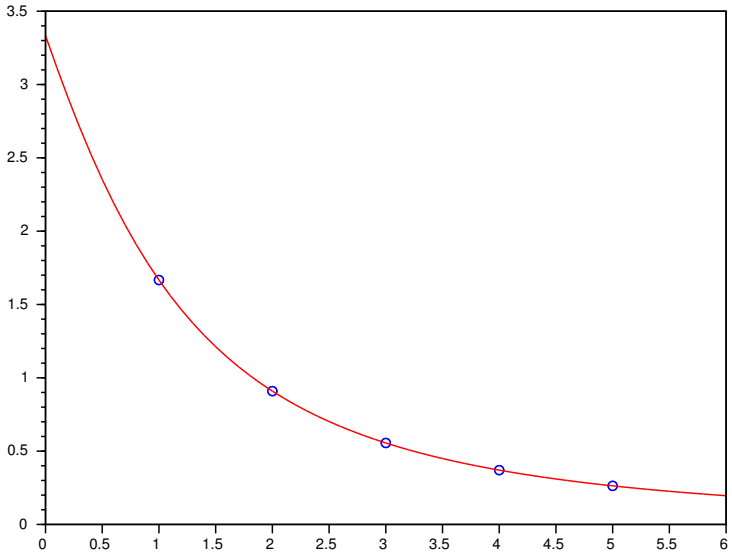


Figura 15.3: Ajuste da curva

Aula 16

Interpolação

16.1 Interpolação

O problema de determinação do **polinômio interpolador** pode ser enunciado da seguinte forma: dados $n+1$ pontos, $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ obter o polinômio $p_n(x)$ de grau n que passe por estes pontos.

Seja

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

Os coeficientes $a_0, a_1, \ldots, a_n$ podem ser obtidos resolvendo o seguinte sistema de equações lineares:

$$\begin{aligned}
p_n(x_0) &= a_n x_0^n + a_{n-1} x_0^{n-1} + \cdots + a_1 x_0 + a_0 = y_0 \\
p_n(x_1) &= a_n x_1^n + a_{n-1} x_1^{n-1} + \cdots + a_1 x_1 + a_0 = y_1 \\
&\vdots \\
p_n(x_n) &= a_n x_0^n + a_{n-1} x_n^{n-1} + \cdots + a_1 x_n + a_0 = y_n
\end{aligned}$$

que pode ser escrito na forma matricial como

$$\begin{pmatrix} x_0^n & x_0^{n-1} & \cdots & 1 \\ x_1^n & x_1^{n-1} & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ x_n^n & x_n^{n-1} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Contudo, este método não é eficiente do ponto de vista numérico.

16.2 Método de Lagrange

O polinômio interpolador $p_n(x)$ pode ser expresso como uma combinação linear de polinômios $L_i(x)$ de grau n:

$$p_n(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x)$$

onde

$$L_i(x) = \frac{(x-x_0)}{(x_i-x_0)} \frac{(x-x_1)}{(x_i-x_1)} \cdots \frac{(x-x_{i-1})}{(x_i-x_{i-1})} \frac{(x-x_{i+1})}{(x_i-x_{i+1})} \cdots \frac{(x-x_{n-1})}{(x_i-x_{n-1})} \frac{(x-x_n)}{(x_i-x_n)}$$

O numerador de $L_i(x)$ é formado pelos fatores $(x - x_k)$ onde $k = 0,1,2,\ldots,n$ excetuando $k = i$. O denominador de $L_i(x)$ é formado pelos fatores $(x_i - x_k)$ onde $k = 0,1,2,\ldots,n$ excetuando $k = i$.

Note que pela definição $L_i(x_k)$ é igual a **um** se $i = k$ e é igual a **zero** se $i \neq k$.

Se os pontos foram originalmente gerados por uma função analítica $f(x)$, para qualquer ponto no intervalo (x_0, x_n) o erro do polinômio interpolador é

$$f(x) - p_n(x) = \frac{f^{(n+1)}(t(x))}{n+1!} \prod_{i=0}^{n} (x - x_i)$$

onde $t(x)$ é desconhecido pas satisfaz $x_0 < t(x) < x_n$.

16.2.1 Exemplo

Encontre o polinômio que interpola os pontos a seguir:

x	1,1	1,2	1,4
y	0,3	0,2	0,15

$$p_2(x) = 0{,}3L_0(x) + 0{,}2L_1(x) + 0{,}15L_2(x)$$

onde

$$L_0(x) = \frac{(x - 1{,}2)}{(1{,}1 - 1{,}2)} \frac{(x - 1{,}4)}{(1{,}1 - 1{,}4)}$$

$$L_1(x) = \frac{(x - 1{,}1)}{(1{,}2 - 1{,}1)} \frac{(x - 1{,}4)}{(1{,}2 - 1{,}4)}$$

$$L_2(x) = \frac{(x - 1{,}1)}{(1{,}4 - 1{,}1)} \frac{(x - 1{,}2)}{(1{,}4 - 1{,}2)}$$

16.3 Interpolação linear segmentada

Considere que os pontos estão organizados de maneira crescente de forma que $x_{k+1} > x_k$. Podemos obter um polinômio de primeira ordem (uma reta) que interpola os pontos no intervalo $[x_k, x_{k+1}]$. Este polinômio é

$$P_i(x) = y_i \frac{(x - x_{i+1})}{(x_i - x_{i+1})} + y_{i+1} \frac{(x - x_i)}{(x_{i+1} - x_i)}$$

Para cada intervalo podemos obter uma reta que interpola os pontos, como pode ser visto na Figura 16.1.

16.4 Interpolação cúbica segmentada

Uma desvantagem da interpolação linear segmentada é que a função resultante não é suave, e sua derivada não é contínua. Um dos tipos de interpolação segmentada mais utilizados é a interpolação cúbica segmentada, que em cada intervalo é definida por um polinômio de ordem 3. Desta maneira é possível garantir que a primeira e a segunda derivada sejam contínuas.

O polinômio interpolador $s(x)$ possui algumas propriedades

- em cada intervalo $[x_k, x_{k+1}]$ a função $s(x)$ é um polinômio cúbico;
- a função $s(x)$ interpola os pontos (x_k, y_k);
- a primeira e a segunda derivada de $s(x)$ são contínuas.

Da primeira hipótese podemos definir o polinômio $s_i(x)$ no intervalo i entre x_i e x_{i+1} como:

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

Da segunda hipótese temos que

$$s_i(x_i) = y_i \quad e \quad s_i(x_{i+1}) = y_{i+1} \quad i = 0{,}1{,}2, \ldots, n-1$$

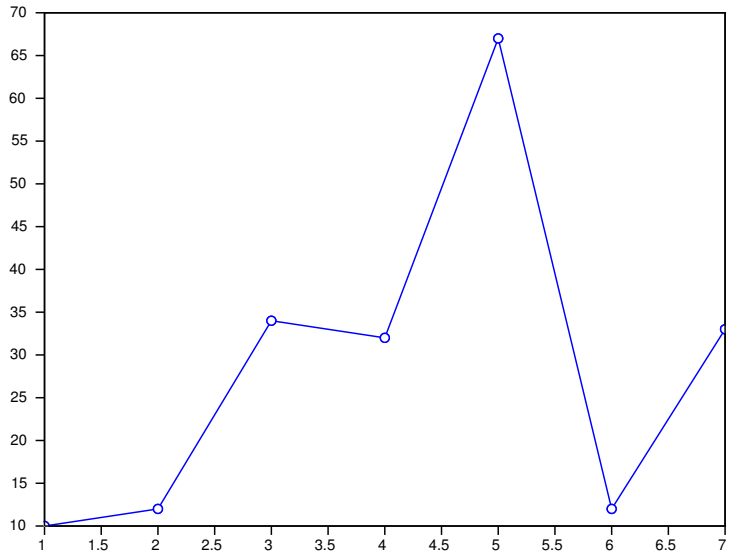


Figura 16.1: Interpolação Linear Segmentada

Da terceira hipótese temos que

$$s_i'(x_{i+1}) = s_{i+1}'(x_{i+1}) \quad e \quad s_i''(x_{i+1}) = s_{i+1}''(x_{i+1}) \quad i = 0,1,2,\ldots,n-2$$

Como temos n intervalos e para cada intervalo temos que determinar 4 parâmetros (a_i, b_i, c_i, d_i), precisamos de $4n$ equações para montar o sistema de equações que definirá os parâmetros. Nas hipóteses acima foram definidas $4n-2$ equações, e para completar o conjunto precisamos definir mais duas. Uma das possibilidades consiste em escolher

$$s_0''(x_0) = 0 \quad e \quad s_{n-1}''(x_n) = 0$$

Definidas as $4n$ equações podemos desenvolver o sistema de equações. Para tanto vamos fazer a seguinte mudança de variáveis $h_i = x_{i+1} - x_i$.

O sistema de equações se torna:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ h_0 & 2h_0+2h_1 & h_1 & 0 & 0 & \cdots & 0 \\ 0 & h_1 & 2h_1+2h_2 & h_2 & 0 & \cdots & 0 \\ 0 & 0 & h_2 & 2h_2+2h_3 & h_3 & & 0 \\ \vdots & \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & h_{n-3} & 2h_{n-3}+2h_{n-2} & h_{n-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ 3\frac{y_2-y_1}{h_1} - 3\frac{y_1-y_0}{h_0} \\ 3\frac{y_3-y_2}{h_2} - 3\frac{y_2-y_1}{h_1} \\ \vdots \\ 3\frac{y_{n-1}-y_{n-2}}{h_{n-2}} - 3\frac{y_{n-2}-y_{n-3}}{h_{n-3}} \\ 0 \end{pmatrix} \tag{16.1}$$

Para encontrar os outros parâmetros usamos as seguintes expressões:

$$a_i = y_i, \quad b_i = \frac{3y_{i+1} - 3y_i - 2c_ih_i^2 - c_{i+1}h_i^2}{3h_i} \quad e \quad d_i = \frac{c_{i+1} - c_i}{3h_i}.$$

Parte V

Derivadas e Integrais

Aula 17

Derivação Numérica

17.1 Derivação Numérica

A derivada de uma função pode ser definida como

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Se não está disponível a expressão de $f(x)$, podemos aproximar a derivada por

$$\frac{df(x)}{dx} \approx D_h^+ f(x) = \frac{f(x+h) - f(x)}{h}$$

utilizando um h pequeno. Na expressão acima, chamamos a aproximação de **diferenças progressivas** por utilizar valor "para frente" de x para aproximar a derivada.

Outra possibilidade para aproximar a derivada é

$$\frac{df(x)}{dx} \approx D_h^- f(x) = \frac{f(x-h) - f(x)}{-h} = \frac{f(x) - f(x-h)}{h}$$

utilizando um h pequeno. Na expressão acima, chamamos a aproximação de **diferenças regressivas** por utilizar valor "para trás" de x para aproximar a derivada.

Uma terceira aproximação é chamada de **diferenças centrais**, e é gerada pela média das duas aproximações anteriores.

$$\frac{df(x)}{dx} \approx D_h^0 f(x) = \frac{D_h^+ f(x) + D_h^- f(x)}{2} = \frac{f(x+h) - f(x-h)}{2h}$$

17.2 Erros

Vamos calcular o erro que cada aproximação comete ao aproximar a derivada.

O erro da aproximação das **diferenças progressivas** é

$$D_h^+ f(x) - f'(x) = \frac{f(x+h) - f(x)}{h} - f'(x)$$

Utilizando a série de Taylor temos que $f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \mathcal{O}(h^3)$. E portanto

$$\begin{aligned} D_h^+ f(x) - f'(x) &= \frac{\left(f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \mathcal{O}(h^3)\right) - f(x)}{h} - f'(x) \\ &= \frac{h}{2} f''(x) + \frac{\mathcal{O}(h^3)}{h} \end{aligned}$$

Observe que a primeira parcela da expressão acima é proporcional a h, e portanto espera-se que ao utilizar h com a metade do tamanho, o erro também seja diminuído pela metade.

Vamos agora calcular o erro para a aproximação das **diferenças regressivas**

$$D_h^- f(x) - f'(x) = \frac{f(x) - f(x-h)}{h} - f'(x)$$

Utilizando a série de Taylor temos que $f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3)$. E portanto

$$\begin{aligned} D_h^- f(x) - f'(x) &= \frac{f(x) - \left(f(x) - hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3)\right)}{h} - f'(x) \\ &= -\frac{h}{2}f''(x) + \frac{\mathcal{O}(h^3)}{h} \end{aligned}$$

Observe que novamente a primeira parcela da expressão acima é proporcional a h, e portanto espera-se que ao utilizar h com a metade do tamanho, o erro também seja diminuído pela metade.

Já o erro da aproximação das **diferenças centrais** é

$$\begin{aligned} D_h^0 f(x) - f'(x) &= \frac{\left(f(x) + hf'(x) + \frac{h^2}{2}f''(x) + O(h^3)\right) - \left(f(x) - hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3)\right)}{2h} - f'(x) \\ &= \frac{\mathcal{O}(h^3)}{h} \end{aligned}$$

Portanto o erro envolve termos cúbicos da série de Taylor, divididos por h, ou seja, o erro é função quadrática de h e portanto ao dividir h por dois, o erro será aproximadamente um quarto.

17.3 Exemplo

Vamos calcular aproximações para a derivada de $sen(x)$ para $x = 1$, utilizando as três aproximações apresentadas, com diferentes valores de h.

Tabela 17.1: Exemplo

	h=0.1	h=0.01	h=0.001
$D_h^+ f(x)$	0.497363752535	0.536085981012	0.539881480360
$D_h^+ f(x) - f'(x)$	0.042938553333	0.004216324856	0.000420825508
$D_h^- f(x)$	0.581440751804	0.544500620738	0.540722951275
$D_h^- f(x) - f'(x)$	- 0.041138445936	- 0.004198314869	- 0.000420645407
$D_h^0 f(x)$	0.539402252170	0.540293300875	0.540302215818
$D_h^0 f(x) - f'(x)$	0.000900053698	0.000009004993	0.000000090050

Podemos observar várias propriedades neste exemplo:

- A ordem de grandeza do erro é a mesma, utilizando as diferenças progressivas e regressivas.
- A ordem de grandeza do erro das diferenças centrais é menor que das outras técnicas.
- Utilizando as diferenças progressivas e regressivas o erro caí aproximadamente 10 vezes ao reduzir h 10 vezes.
- Utilizando as diferenças centrais o erro caí aproximadamente 100 vezes ao reduzir h 10 vezes.

17.4 Escolha do intervalo de derivação

Todas as aproximações apresentadas dependem da escolha do intervalo de derivação h. Foi mostrado que o erro depende de h e que quando menor h menor será o erro. Contudo, precisamos levar em conta que as operações serão realizadas em uma máquina (computador ou calculadora) que possui precisão finita para armazenar os números, e devemos tomar cuidado com o efeito conhecido como cancelamento catastrófico.

Como exemplo, a aproximação das diferenças progressivas é

$$\frac{f(x+h)-f(x)}{h}.$$

Se h possui um valor muito pequeno então ocorre cancelamento catastrófico no numerador da expressão. Portanto devemos escolher h pequeno, mas que não seja demasiadamente pequeno.

Considere um erro ε ao calcular o valor da função. Logo, usando diferenças progressivas temos que

$$D_h^+ f(x) = \frac{f(x+h)+\varepsilon - f(x)+\varepsilon}{h}.$$

e o erro de **arredondamento** é dado por

$$\frac{2\varepsilon}{h}$$

e quanto menor h maior será o erro.

Vamos analisar o erro da aproximação das diferenças progressivas da derivada de $sen(x)$ para diferentes valores de h.

Tabela 17.2: Escolha do intervalo de derivação

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Erro	0.042938553333	0.004216324856	0.000420825508	0.000042074450	0.000004207362
h	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
Erro	0.000000420747	0.000000041828	0.000000014072	- 0.000000052541	0.000000058481
h	10^{-11}	10^{-22}	10^{-13}	10^{-14}	10^{-15}
Erro	0.000001168704	- 0.000043240217	0.000733915900	0.007395254048	- 0.014809206444

Observe que o erro diminui ao diminuir o valor de h até aproximadamente $h = 10^{-8}$. Se diminuímos mais ainda o valor de h o efeito do cancelamento catastrófico fica evidente pois o erro aumenta ao diminuir o valor de h.

17.5 Exercício

Exercise 40 Calcule as seguintes derivadas utilizando todas as técnicas com $h = 0.01$ e $h = 0.001$

- Derivada primeira de $f(x)$ onde $f(x) = sin(x)$ e $x = 2$
- Derivada primeira de $f(x)$ onde $f(x) = e^x$ e $x = 1$
- Derivada segunda de $f(x)$ onde $f(x) = e^x$ e $x = 1$
- Derivada primeira de $f(x)$ onde $f(x) = sin(cos(e^x + 1))$ e $x = 1$
- Derivada segunda de $f(x)$ onde $f(x) = sin(cos(x) + x + x^2)$ e $x = 0.1$

Exercise 41 As tensões na entrada, v_i, e saída, v_o, de um amplificador foram medidas em regime estacionário conforme tabela abaixo.

0.	0.5	1.	1.5	2.	2.5	3.	3.5	4.	4.5	5.
0.	1.05	1.83	2.69	3.83	4.56	5.49	6.56	6.11	7.06	8.29

onde a primeira linha é a tensão de entrada em volts e a segunda linha é tensão de saída em volts. Sabendo que o ganho é definido como

$$\frac{\partial v_o}{\partial v_i}.$$

Calcule o ganho quando $v_i = 1$, $v_i = 0$, $v_i = 4.5$ e $v_i = 5$.

Aula 18

Outros tipos de derivadas

18.1 Fórmula Genérica

A derivada de primeira ordem pode ser aproximada usando 2 pontos como:

$$f'(x) \approx K_1 f(x+h_1) + K_2 f(x+h_2)$$

Por exemplo:

- Diferenças Progressivas: $K_1 = 1/h$, $K_2 = -1/h$, $h_1 = h$ e $h_2 = 0$.
- Diferenças Regressivas: $K_1 = 1/h$, $K_2 = -1/h$, $h_1 = 0$ e $h_2 = -h$.
- Diferenças Centrais: $K_1 = 1/(2h)$, $K_2 = -1/(2h)$, $h_1 = h$ e $h_2 = -h$.

Observe que usando a Série de Taylor podemos escolher os valores ótimos da regra de dois pontos.

$$\begin{aligned} f'(x) &\approx K_1 f(x+h_1) + K_2 f(x+h_2) && (18.1)\\ &\approx K_1 \left(f(x) + h_1 f'(x) + \frac{h_1^2}{2} f''(x) + \mathcal{O}(h_1^3) \right) && (18.2)\\ &\quad + K_2 \left(f(x) + h_2 f'(x) + \frac{h_2^2}{2} f''(x) + \mathcal{O}(h_2^3) \right) && (18.3) \end{aligned}$$

Para eliminar o primeiro termo temos que fazer $K_1 + K_2 = 0$.

Para fazer o segundo termo ser igual a derivada, temos que fazer $K_1 h_1 + K_2 h_2 = 1$.

Para eliminar o terceiro termo temos que fazer $K_1 \frac{h_1^2}{2} + K_2 \frac{h_2^2}{2} = 0$.

Para eliminar o último termo temos que fazer $h_1 = h_2 = 0$, o que não é possível, portanto temos que manter os valores h pequenos.

Note que a regra "diferenças centrais" obedece todas as equações acima.

A derivada de primeira ordem pode ser aproximada usando 3 pontos como:

$$f'(x) \approx K_1 f(x+h_1) + K_2 f(x+h_2) + K_3 f(x+h_3)$$

Uma possibilidade é criar uma regra progressiva:

$$f'(x) \approx K_1 f(x) + K_2 f(x+h) + K_3 f(x+2h)$$

Usando a Série de Taylor podemos escolher os valores ótimos da regra de três pontos.

$$\begin{aligned} f'(x) &\approx K_1 f(x) + K_2 f(x+h) + K_3 f(x+2h) && (18.4)\\ &\approx K_1 f(x) && (18.5)\\ &\quad + K_2 \left(f(x) + h f'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(x) + \mathcal{O}(h^4) \right) && (18.6)\\ &\quad + K_3 \left(f(x) + 2h f'(x) + \frac{(2h)^2}{2} f''(x) + \frac{(2h)^3}{6} f'''(x) + \mathcal{O}((2h)^2) \right) && (18.7) \end{aligned}$$

Portanto

$$K_1 + K_2 + K_3 = 0 \tag{18.8}$$
$$K_2 h + K_3 2h = 1 \tag{18.9}$$
$$K_2 \frac{h^2}{2} + K_3 4\frac{h^2}{2} = 0 \tag{18.10}$$
$$\tag{18.11}$$

Logo, $K_1 = -3/(2h)$, $K_2 = 2/h$ e $K_3 = -1/(2h)$.

Nesse caso, o erro será dado por $\mathcal{O}(h^2)$ (PROVE!), que é melhor que a regra com apenas dois pontos.

18.2 Segunda derivada

Vamos agora obter uma aproximação para a segunda derivada da função. Observe que

$$f''(x) = \frac{d}{dx}\left(\frac{d}{dx} f(x)\right).$$

Vamos agora utilizar a aproximação das diferenças centrais para aproximar a derivada da função:

$$f''(x) \approx D^2 f(x) = \frac{f'(x+\delta) - f'(x-\delta)}{2\delta}.$$

Podemos utilizar a mesma aproximação novamente:

$$f''(x) \approx D^2 f(x) = \frac{\left(\frac{f(x+\delta+\delta)-f(x+\delta-\delta)}{2\delta}\right) - \left(\frac{f(x-\delta+\delta)-f(x-\delta-\delta)}{2h}\right)}{2\delta}.$$

$$D^2 f(x) = \frac{f(x+2\delta) - 2f(x) + f(x-2\delta)}{(2\delta)^2}.$$

Fazendo a mudança de variáveis $h = 2\delta$

$$D^2 f(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Para calcular o erro vamos utilizar novamente a série de Taylor:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(x) + \mathcal{O}(h^4)$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{6} f'''(x) + \mathcal{O}(h^4)$$

e portanto

$$D^2 f(x) - f''(x) = \frac{2\mathcal{O}(h^4)}{h^2} = \mathcal{O}(h^2)$$

Logo espera-se que ao utilizar h com a metade do tamanho, o erro seja reduzido 4 vezes.

Podemos criar regras de 3, 4, 5 pontos ou mais para estimar a segunda derivada. De maneira semelhante podemos criar regras para estimar derivadas superiores.

18.3 Gradiente

Seja $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ uma função analítica em um domínio $\mathcal{D} \subseteq \mathbb{R}^n$. O gradiente desta função é um vetor coluna denotado

$$\nabla f(x) = \frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

18.4 Jacobiano

Seja $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ uma função analítica em um domínio $\mathcal{D} \subseteq \mathbb{R}^n$. O Jacobiano desta função é a generalização do gradiente para funções vetoriais. O Jacobiano é um vetor denotado

$$J = \left[\begin{array}{cccc} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{array} \right] = \left[\begin{array}{cccc} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{array} \right]$$

Usualmente é utilizada a aproximação por **diferenças centrais** para estimar cada uma das $m \times n$ derivadas.

Observe que o **gradiente** é igual ao **jacobiano** transposto, e portanto podemos usar a função `jacobiano` para estimar o gradiente.

18.5 Hessiana

A Hessiana de uma função $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ é o equivalente da segunda derivada para funções com várias variáveis e é uma matriz simétrica denotada

$$\nabla^2 f(x) = \frac{\partial^2 f(x)}{\partial x^2} = \frac{\partial \left(\frac{\partial f(x)}{\partial x} \right)}{\partial x} = \left[\begin{array}{cccc} \frac{\partial^2 f}{\partial x_1 x_1} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n x_n} \end{array} \right]$$

Observe que a **hessiana** é igual ao **jacobiano** do **gradiente**, e portanto podemos usar duas vezes a função `jacobiano` para estimar a **hessiana**.

Aula 19

Integral Numérica - Newton-Cotes

O objetivo das técnicas é aproximar a integral de uma função $f(x)$ definida em um intervalo $[a, b]$:

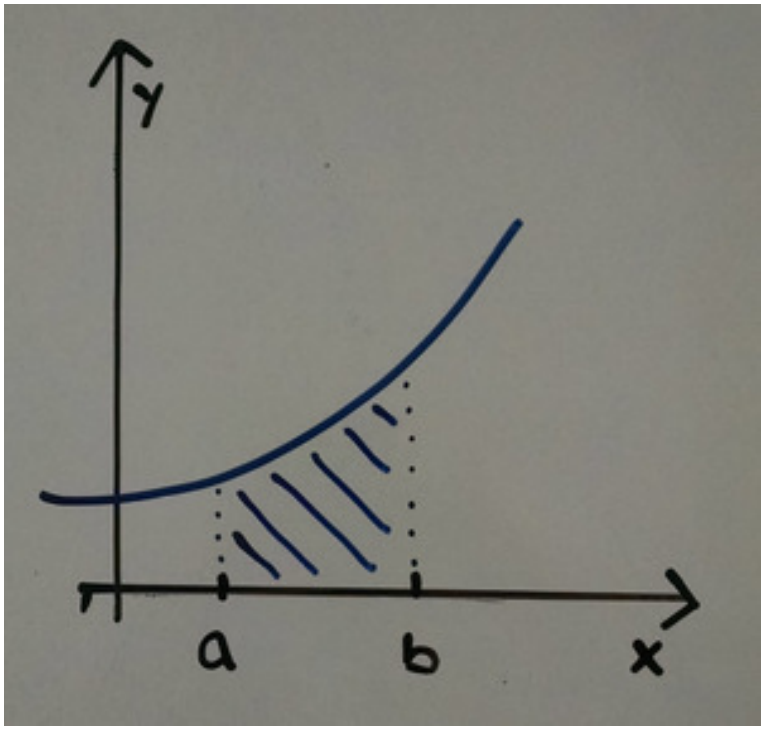


Figura 19.1: Função $f(x)$.

19.1 Regra do Ponto Médio

Idéia: no intervalo $[a, b]$ a função $f(x)$ pode ser aproximada por $f(x_m)$ onde $x_m = \frac{a+b}{2}$.

$$\int_a^b f(x)dx \approx \int_a^b f(x_m)dx = (b-a)f(x_m)$$

19.1.1 Cáculo do erro

Usando a série de Taylor

$$f(x) = f(x_m) + f'(x_m)(x - x_m) + \frac{f''(t(x))}{2}(x - x_m)^2$$

onde p é um ponto desconhecido tal que $a < t(x) < b$.

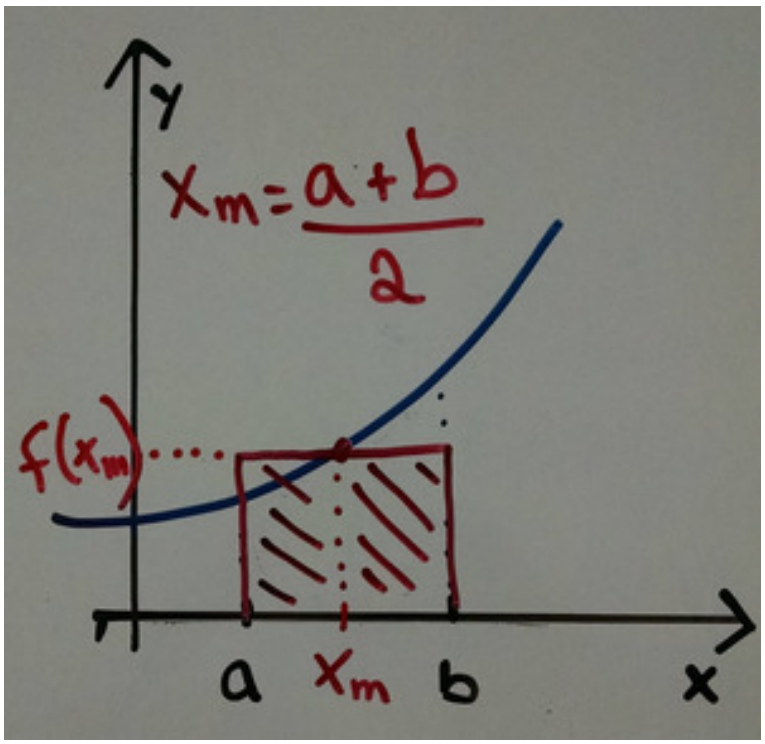


Figura 19.2: Regra do Ponto Médio.

$$
\begin{aligned}
Erro &= \int_a^b f(x)dx - \int_a^b f(x_m)dx && (19.1)\\
&= \int_a^b \left(f(x_m) + f'(x_m)(x-x_m) + \frac{f''(t(x))}{2}(x-x_m)^2 \right) dx - \int_a^b f(x_m)dx && (19.2)\\
&= \int_a^b f(x_m)dx + \int_a^b f'(x_m)(x-x_m)dx + \int_a^b \frac{f''(t(x))}{2}(x-x_m)^2dx - \int_a^b f(x_m)dx && (19.3)\\
&= \int_a^b f'(x_m)(x-x_m)dx + \int_a^b \frac{f''(t(x))}{2}(x-x_m)^2dx && (19.4)
\end{aligned}
$$

Observe que $f'(x_m)$ é uma constante. Além disso, note que $a < t(x) < b$ e que $(x-x_m)^2 > 0$ e portanto $\exists p \in (a,b)$ tal que

$$
\begin{aligned}
Erro &= f'(x_m)\int_a^b (x-x_m)dx + \frac{f''(p)}{2}\int_a^b (x-x_m)^2dx && (19.5)\\
&= f'(x_m)\left[\frac{(x-x_m)^2}{2}\right]_a^b dx + \frac{f''(p)}{2}\left[\frac{(x-x_m)^3}{3}\right]_a^b dx && (19.6)\\
&= 0 + \frac{f''(p)}{2}\left(\frac{(b-x_m)^3}{3} - \frac{(a-x_m)^3}{3}\right) && (19.7)\\
Erro &= \frac{(b-a)^3}{24}f''(p) && (19.8)
\end{aligned}
$$

onde $p \in (a,b)$.

19.1.2 Regra Composta

$$\int_a^b f(x)dx = \int_a^{a+h} f(x)dx + \int_{a+h}^{a+2*h} f(x)dx + \cdots + \int_{b-h}^{b} f(x)dx$$

onde

$$h = \frac{(b-a)}{N}$$

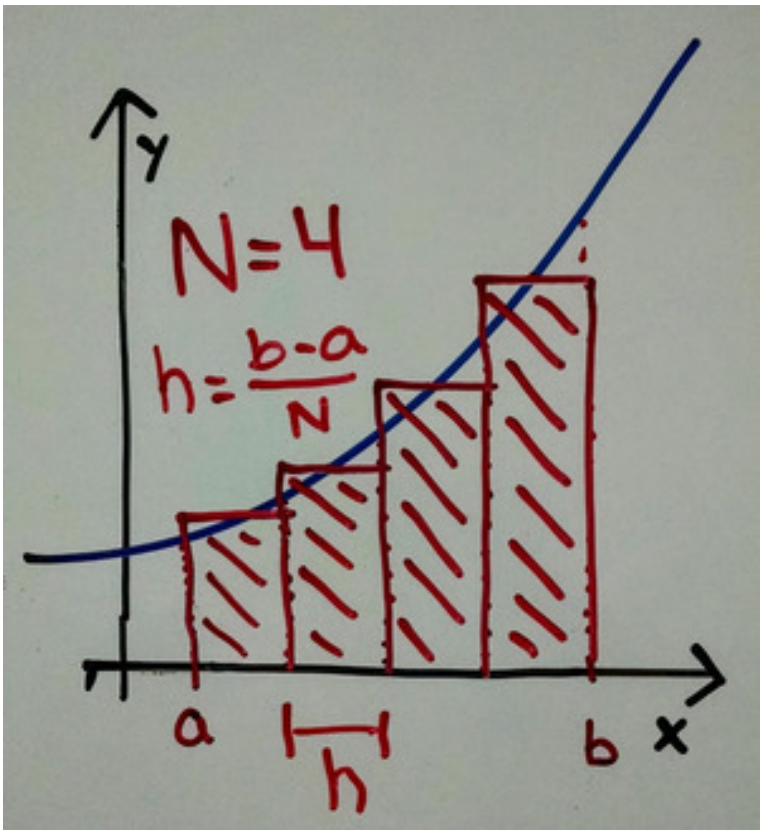


Figura 19.3: Regra do Ponto Médio Composta

Utilizando a regra do ponto médio

$$\int_a^b f(x)dx \approx (a+h-a)f(\frac{a+a+h}{2}) + (a+2h-a-h)f(\frac{2a+3h}{2}) + \cdots + (b-b+h)f(\frac{2b-h}{2})$$

$$\int_a^b f(x)dx \approx hf\left(a+\frac{h}{2}\right) + hf\left(a+\frac{3h}{2}\right) + \cdots + hf\left(b-\frac{h}{2}\right)$$

ou seja

$$\int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{i=1}^{N} f\left(a+\frac{h}{2}(2i-1)\right)$$

O Erro da regra composta será

$$Erro = N\frac{h^3}{24}f''(p) = \frac{(b-a)^3}{24N^2}f''(p)$$

19.2 Regra do Trapézio

Ideia: no intervalo $[a,b]$ a função $f(x)$ pode ser aproximada pelo polinômio interpolados de Lagrange de primeira ordem.

$$f(x) \approx f(b)\frac{x-a}{b-a} - f(a)\frac{x-b}{b-a}$$

$$\int_a^b f(x)dx \approx \int_a^b \left(f(b)\frac{x-a}{b-a} - f(a)\frac{x-b}{b-a}\right)dx = (b-a)\left(\frac{f(a)+f(b)}{2}\right)$$

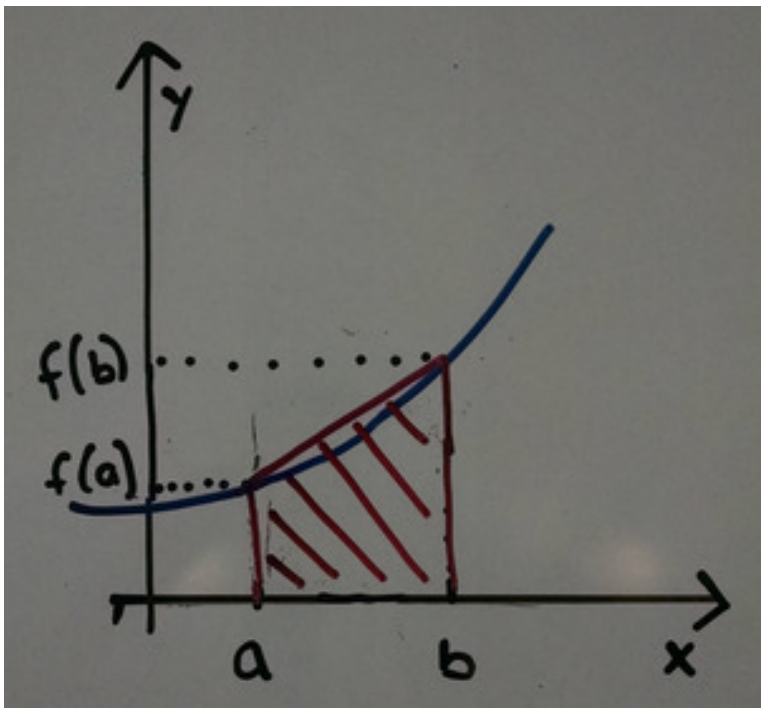


Figura 19.4: Regra do Trapézio.

19.2.1 Cálculo do erro

Os polinômios interpoladores de Lagrange podem ser escritos como

$$f(x) = f(b)\frac{x-a}{b-a} - f(a)\frac{x-b}{b-a} + E(x)$$

onde $E(x)$ é o termo do erro:

$$E(x) = \frac{f''(t(x))}{2}(x-b)(x-a)$$

onde t(x) é desconhecido mas sabe-se que $a < t(x) < b$.

Portanto

$$\begin{aligned} Erro &= \int_a^b f(x)dx - (b-a)\left(\frac{f(a)+f(b)}{2}\right) && (19.9)\\ &= \int_a^b E(x)dx && (19.10)\\ &= \int_a^b \frac{f''(t(x))}{2}(x-b)(x-a)dx && (19.11)\\ & && (19.12) \end{aligned}$$

Observe que $a < t(x) < b$ e que $(x-b)(x-a) < 0$ e portanto $\exists p \in (a,b)$ tal que

$$\begin{aligned} Erro &= \frac{f''(p)}{2}\int_a^b (x-b)(x-a)dx && (19.13)\\ &= -\frac{(b-a)^3}{12}f''(p) && (19.14) \end{aligned}$$

onde $p \in (a,b)$.

19.2.2 Regra Composta

Considere agora a integral

$$\int_a^b f(x)dx = \int_a^{a+h} f(x)dx + \int_{a+h}^{a+2*h} f(x)dx + \cdots + \int_{b-h}^{b} f(x)dx$$

onde

$$h = \frac{(b-a)}{N}$$

Utilizando a regra do trapézio

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{N} \frac{f(a+h(i-1)) + f(a+hi)}{2} \tag{19.15}$$

$$\approx h\left(\frac{f(a)}{2} + \frac{f(b)}{2} + \sum_{i=1}^{N-1} f(a+hi)\right) \tag{19.16}$$

O Erro da regra composta será

$$Erro = -N\frac{h^3}{12}f''(p) = -\frac{(b-a)^3}{12N^2}f''(p)$$

onde $a < p < b$.

19.3 Regra de Simpson

Idéia: no intervalo $[a, b]$ a função $f(x)$ pode ser aproximada pelo polinômio interpolador de Lagrange de segunda ordem. Seja m o ponto médio $m = (a + b)/2$):

$$f(x) \approx f(a)\frac{(x-m)(x-b)}{(a-m)(a-b)} + f(m)\frac{(x-a)(x-b)}{(m-a)(m-b)} + f(b)\frac{(x-a)(x-m)}{(b-a)(b-m)}$$

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a) + 4f(m) + f(b)\right)$$

19.3.1 Cáculo do erro

$$Erro = -\frac{(b-a)^5}{2880}\left.\frac{d^4y(x)}{dx^4}\right|_{x=p} \tag{19.17}$$

onde $p \in (a, b)$.

19.3.2 Regra Composta

Considere agora a integral

$$\int_a^b f(x)dx = \int_a^{a+h} f(x)dx + \int_{a+h}^{a+2*h} f(x)dx + \cdots + \int_{b-h}^{b} f(x)dx$$

onde

$$h = \frac{(b-a)}{N}$$

Utilizando a regra de Simpson

$$\int_a^b f(x)dx \approx \frac{h}{6}\sum_{i=1}^{N}\left(f(a+h(i-1)) + 4f(a+h(i-1/2)) + f(a+hi)\right) \tag{19.18}$$

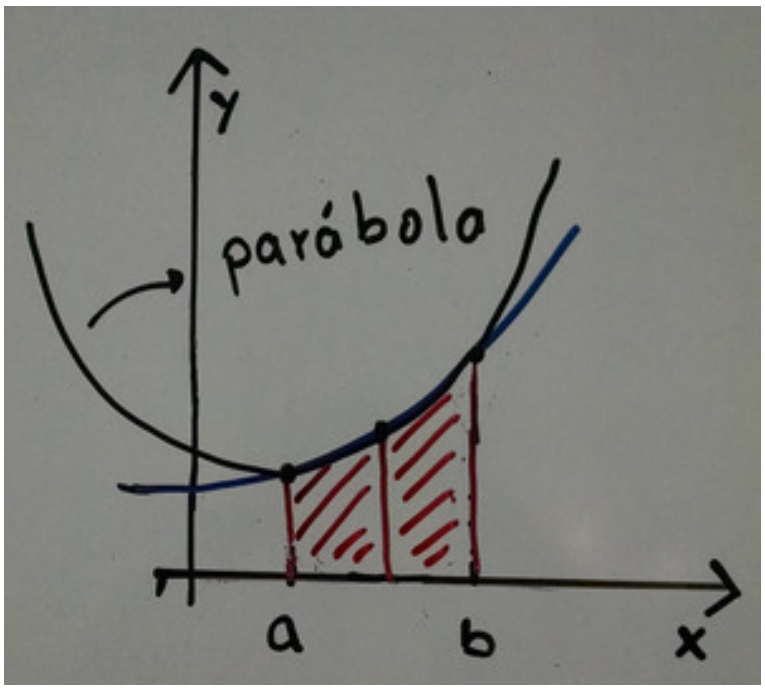


Figura 19.5: Regra de Simpson.

ou seja

$$\int_a^b f(x)dx \approx \frac{h}{6}\left(f(a)+f(b)+4\sum_{i=1}^{N} f(a+h(i-1/2))+2\sum_{i=1}^{N-1} f(a+hi)\right)$$

O Erro da regra composta será

$$Erro = -N\frac{(h)^5}{2880}\left.\frac{d^4y(x)}{dx^4}\right|_{x=p} = -\frac{(b-a)^5}{2880N^4}\left.\frac{d^4y(x)}{dx^4}\right|_{x=p}$$

onde $a < p < d$.

19.4 Exercícios

Exercise 42 Aproxime as integrais utilizando todas as regras simples e todas as regras compostas com $N = 2$. Estime também o valor máximo do erro.

- $\int_0^1 xdx$
- $\int_0^1 x^3dx$
- $\int_0^1 x^5dx$

Aula 20

Quadratura Gaussiana

20.1 Integral exata de polinômios

Seja $f(x)$ uma função polinomial definida no intervalo $[-1,1]$ da qual se deseja conhecer a integral:

$$\int_{-1}^{1} f(x)dx$$

Esta integral será aproximada pela seguinte regra:

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{N} A_i f(x_i)$$

Onde x_i são N valores escolhidos arbitrariamente no intervalo $[-1,1]$. Gostaríamos de obter os valores de A_i que melhor aproximam a integral acima. Como exemplo, vamos definir que $N = 3$, ou seja, vamos medir o valor da função em 3 pontos distintos. Mais adiante vamos generalizar para o caso com mais pontos. Neste caso

$$\int_{-1}^{1} f(x)dx \approx A_1 f(x_1) + A_2 f(x_2) + A_3 f(x_3)$$

Gostaríamos de garantir que a integral seja exata, caso a função $f(x)$ seja uma função constante $f(x) = 1$. Neste caso

$$f(x) = 1 \quad \text{e} \quad \int_{-1}^{1} 1dx = 2, \quad \text{portanto} \quad A_1 f(x_1) + A_2 f(x_2) + A_3 f(x_3) = 2,$$

logo, $A_1 1 + A_2 1 + A_3 1 = 2$.

Gostaríamos também de garantir que a integral seja exata, caso a função $f(x)$ seja uma reta $f(x) = x$. Neste caso

$$f(x) = x \quad \text{e} \quad \int_{-1}^{1} xdx = 0, \quad \text{portanto} \quad A_1 f(x_1) + A_2 f(x_2) + A_3 f(x_3) = 0,$$

logo, $A_1 x_1 + A_2 x_2 + A_3 x_3 = 0$.

Também gostaríamos de garantir que a integral seja exata, caso a função $f(x)$ seja uma parábola $f(x) = x^2$. Neste caso

$$f(x) = x^2 \quad \text{e} \quad \int_{-1}^{1} x^2 dx = 2/3, \quad \text{portanto} \quad A_1 f(x_1) + A_2 f(x_2) + A_3 f(x_3) = 2/3$$

logo, $A_1 x_1^2 + A_2 x_2^2 + A_3 x_3^2 = 2/3$.

Portanto, para quaisquer três pontos x_1, x_2 e x_3 que forem escolhidos no intervalo $[-1,1]$, se as constantes A_1, A_2 e A_3 satisfizeram

$$\begin{cases} A_1 1 + A_2 1 + A_3 1 = 2 \\ A_1 x_1 + A_2 x_2 + A_3 x_3 = 0 \\ A_1 x_1^2 + A_2 x_2^2 + A_3 x_3^2 = 2/3 \end{cases}$$

então a regra proposta vai integral exatamente polinômios de até graus 2^1.

Portanto, medindo a função em N pontos é possível criar uma regra que garante a exatidão no cálculo da integral de polinômios de grau $N-1$. Para tanto basta que:

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_N \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_N^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^{N-1} & x_2^{N-1} & x_3^{N-1} & \dots & x_N^{N-1} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_N \end{pmatrix} = \begin{pmatrix} 2/1 \\ 0 \\ 2/3 \\ \vdots \\ 0 \text{ ou } 2/N \end{pmatrix}$$

20.1.1 Exemplo 1

Gostaríamos de calcular a integral $\int_{-1}^{1} 1 + 2x + 3x^2 dx$ medindo a função nos pontos $x_1 = 0$, $x_2 = 0.5$ e $x_3 = 1$. O valor analítico da integral é 4. Vamos calcular as constantes A_1, A_2 e A_3 conforme apresentado.

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 1 \\ 0 & 0.25 & 1 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} = \begin{pmatrix} 2/1 \\ 0 \\ 2/3 \end{pmatrix}$$

Portanto $A_1 = 10/3$, $A_2 = -8/3$ e $A_3 = 4/3$. A integral será calculada como

$$\frac{10}{3} f(0) - \frac{8}{3} f(0{,}5) + \frac{4}{3} f(1) = \frac{10}{3} 1 - \frac{8}{3} 2{,}75 + \frac{4}{3} 6 = 4$$

E portanto a integral foi calculada com exatidão.

20.1.2 Exemplo 2

Gostaríamos de calcular a integral $\int_{-1}^{1} 1 + 2x + 3x^2 + 4x^3 + 5x^4 dx$ medindo a função nos pontos $x_1 = 0$, $x_2 = 0.5$ e $x_3 = 1$. O valor analítico da integral é 6. Vamos usar as mesmas constantes calculadas anteriormente $A_1 = 10/3$, $A_2 = -8/3$ e $A_3 = 4/3$. A integral será calculada como

$$\frac{10}{3} f(0) - \frac{8}{3} f(0{,}5) + \frac{4}{3} f(1) = \frac{10}{3} 1 - \frac{8}{3} 3.5625 + \frac{4}{3} 15 = 13.8333...$$

E portanto a integral foi calculada com bastante erro. Esta regra foi obtida para garantir a integração exata apenas de parábolas. Como esta função é de grau 5 vão ocorrer erros.

É fácil imaginar que se outros pontos x_i fossem escolhidos, a aproximação seria melhor. Escolhendo os pontos $x_1 = -0.8$, $x_2 = 0$ e $x_3 = 0.8$ e refazendo todos os cálculos, então obteríamos uma aproximação para a integral 6,13333... a qual é muito melhor que a aproximação obtida anteriormente.

20.2 Quadratura de Gauss-Legendre

A Quadratura de Gauss-Legendre, consistem em escolher os N pontos x_i como as raízes do polinômio de Legendre de grau N. Os polinômios de Legendre são da forma:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left[\left(x^2 - 1 \right)^n \right].$$

[1] Pois soma e integral são operadores lineares.

Por exemplo:

$$P_1(x) = x \quad \text{cuja raíz é } 0 \tag{20.1}$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1) \quad \text{cujas raízes são } \pm\sqrt{3}/3 \tag{20.2}$$

$$P_3(x) = \frac{1}{2}x(5x^2 - 3) \quad \text{cujas raízes são } 0, \pm\sqrt{3/5} \tag{20.3}$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \quad \text{cujas raízes são } \pm\frac{\sqrt{35}}{35}\sqrt{15 + \sqrt{120}}, \pm\frac{\sqrt{35}}{35}\sqrt{15 - \sqrt{120}} \tag{20.4}$$

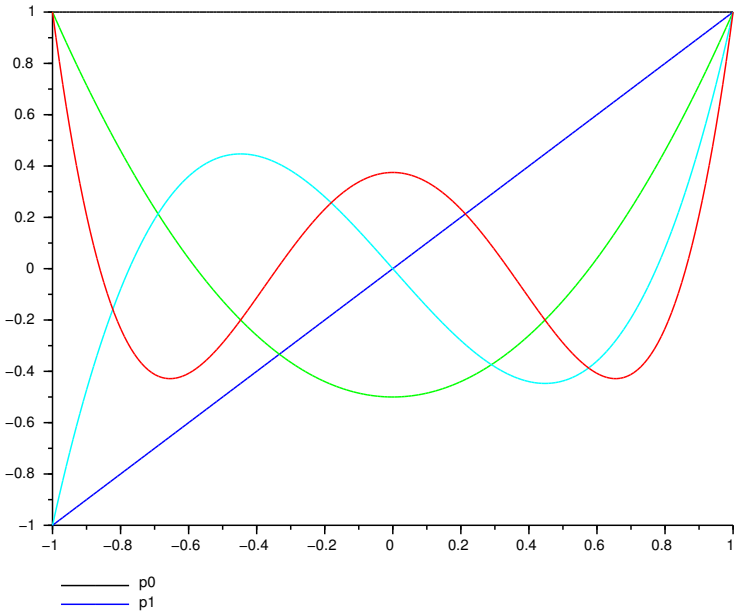


Figura 20.1: Polinômios de Legendre

20.2.1 Exemplo

Gostaríamos de calcular a integral $\int_{-1}^{1} 1 + 2x + 3x^2 + 4x^3 + 5x^4 dx$ medindo a função nos pontos das três raízes do polinômio de Legendre de grau 3: $x_1 = -\sqrt{3/5}$, $x_2 = 0$ e $x_3 = \sqrt{3/5}$. O valor analítico da integral é 6. Utilizando a regra apresentada anteriormente $A_1 = 0.5555...$, $A_2 = 0.8888...$ e $A_3 = 0.5555....$ A integral será aproximada por

$$0.555\overline{5}f(-\sqrt{3/5}) + 0.888\overline{8}f(0) + 0.555\overline{5}f(\sqrt{3/5}) = 6$$

E portanto a integral foi calculada com exatidão. Na sequência vamos mostrar porque as raízes do polinômio de Legendre são uma boa escolha para os pontos de integração.

20.3 Propriedades dos Polinômios de Legendre

Uma importante propriedade dos polinômios de Legendre é

$$\int_{-1}^{1} q(x)P_n(x)dx = 0$$

se $q(x)$ é um polinômio de grau menor que n. Vamos usar esta propriedade para mostrar porque as raízes do polinômio de Legendre são uma boa escolha para os pontos de integração.

Vamos assumir que desejamos encontrar a integral de um polinômio $p_{2n-1}(x)$ de grau $2n-1$. Este polinômio pode ser fatorado como

$$p_{2n-1}(x) = q_{n-1}(x)P_n(x) + r_{n-1}(x)$$

onde $P_n(x)$ é o polinômio de Legendre e $q_{n-1}(x)$ e $r_{n-1}(x)$ são polinômios com grau $n-1$.

A integral de $p_{2n-1}(x)$ é

$$\int_{-1}^{1} p_{2n-1}(x) = \int_{-1}^{1} q_{n-1}(x)P_n(x)dx + \int_{-1}^{1} r_{n-1}(x)dx.$$

Utilizando a propriedade do polinômio de Legendre:

$$\int_{-1}^{1} p_{2n-1}(x) = \int_{-1}^{1} r_{n-1}(x)dx.$$

Ok! Vamos agora aproximar a integral utilizando n pontos distintos:

$$\int_{-1}^{1} p_{2n-1}(x) = \int_{-1}^{1} r_{n-1}(x)dx \approx \sum_{i=1}^{n} A_i p_{2n-1}(x_i)$$

onde os pontos x_i são as raízes do polinômio de Legendre de grau n, da mesma maneira que fizemos no exemplo anterior. Fatorando o polinômio temos que

$$\int_{-1}^{1} p_{2n-1}(x) = \int_{-1}^{1} r_{n-1}(x)dx \approx \sum_{i=1}^{n} A_i p_{2n-1}(x_i) = \sum_{i=1}^{n} A_i q_{n-1}(x)P_n(x_i) + \sum_{i=1}^{n} A_i r_{n-1}(x_i)$$

Observe que no primeiro somatório aparece o termo $P_n(x_i)$. Como os pontos x_i foram escolhidos como as raízes do polinômio de Legendre, então $P_n(x_i) = 0$. Logo,

$$\int_{-1}^{1} p_{2n-1}(x) = \int_{-1}^{1} r_{n-1}(x)dx \approx \sum_{i=1}^{n} A_i r_{n-1}(x_i).$$

Entretanto, $r_{n-1}(x)$ é um polinômio de grau menor que n, e portanto a aproximação acima é exata! Logo, se utilizamos as n raízes do polinômio de Legendre como pontos de integração, então a integral será exata para qualquer polinômio de grau $2n-1$.

20.4 Mudança de variáveis

Até agora todas as integrais foram calculadas para os limites -1 e 1. Caso desejamos utilizar outros limites, podemos fazer a seguinte mudança de variável:

$$t(x) = \frac{b+a}{2} + \frac{b-a}{2}x \tag{20.5}$$

Usando esta mudança de variáveis:

$$\int_a^b f(t)dt = \frac{b-a}{2}\int_{-1}^{1} f(t(x))dx$$

A aproximação da integral então se torna

$$\int_a^b f(t)dt = \frac{b-a}{2}\sum_{i=1}^{N} A_i f(t_i)$$

onde os pontos t_i são calculados usando a mudança de variável (20.5).

20.4.1 Exemplo

Calcule a integral utilizando o método de Gauss-Legendre com 3 pontos:

$$\int_0^1 sen(x)dx$$

Solução:

$$\int_0^1 sen(x)dx = \frac{1-0}{2}\int_{-1}^1 sen(x)dx \approx \frac{1}{2}\sum_{i=1}^{3} A_i f(t_i)$$

As constantes A_i foram calculadas anteriormente: $A_1 = 0.5555...$, $A_2 = 0.8888...$ e $A_3 = 0.5555....$ Os pontos t_i são calculados utilizando a mudança de variáveis acima

$$t_1 = \frac{b+a}{2} + \frac{b-a}{2}x_1 = \frac{1+0}{2} + \frac{1-0}{2}(-\sqrt{3/5})) = 0.112701665379 \quad (20.6)$$

$$t_2 = \frac{b+a}{2} + \frac{b-a}{2}x_1 = \frac{1+0}{2} + \frac{1-0}{2}(0)) = 0.5 \quad (20.7)$$

$$t_3 = \frac{b+a}{2} + \frac{b-a}{2}x_1 = \frac{1+0}{2} + \frac{1-0}{2}(\sqrt{3/5})) = 0.887298334621 \quad (20.8)$$

Portanto a aproximação é

$$\int_0^1 sen(x)dx \approx \frac{1}{2}\sum_{i=1}^{3} A_i f(t_i) = 0.459697930132$$

e a solução analítica é 0.459697694132.

Parte VI

Equações Diferenciais

Aula 21

Solução de problemas de valor inicial - Método de Euler

21.1 Equações diferenciais de primeira ordem

Considere a seguinte equação diferencial de primeira ordem com valor inicial:

$$\frac{dy(t)}{dt} = f(y(t), t) = -2y(t) \tag{21.1}$$
$$y(0) = 1 \tag{21.2}$$

A solução deste problema é

$$y(t) = e^{-2t}$$

21.2 Método de Euler

O método de Euler consiste em utilizar a regra das *diferenças progressivas* para aproximar a derivada presente na equação diferencial. Relembre que

$$\frac{dy(t)}{dt} \approx \frac{y(t+h) - y(t)}{h}$$

Logo, a equação diferencial será aproximada por

$$\frac{dy(t)}{dt} \approx \frac{y(t+h) - y(t)}{h} = f(y(t), t)$$

Isolando o termo $y(t+h)$ temos que

$$y(t+h) = y(t) + hf(y(t), t)$$

21.2.1 Exemplo 1

Considere o problema apresentado anteriormente, em que $f(y(t), t) = -2y(t)$. Para este problema

$$y(t+h) = y(t) + hf(y(t), t) = y(t) - 2hy(t) = (1 - 2h)y(t).$$

Vamos considerar inicialmente que $h = 0.1$

t	Solução analítica	Método de Euler
0	1	1
0,1	$e^{-0.2} = 0{,}8187308$	$(1 - 2 \cdot 0.1)y(0) = 0{,}8 \cdot 1 = 0{,}8$
0,2	$e^{-0.4} = 0{,}6703200$	$(1 - 2 \cdot 0.1)y(0.1) = 0{,}8 \cdot 0.8 = 0{,}64$
0,3	$e^{-0.6} = 0{,}5488116$	$(1 - 2 \cdot 0.1)y(0.2) = 0{,}8 \cdot 0.64 = 0{,}512$
0,4	$e^{-0.8} = 0{,}4493290$	$(1 - 2 \cdot 0.1)y(0.3) = 0{,}8 \cdot 0.512 = 0{,}4096$

21.2.2 Exemplo 2

Considere a seguinte equação diferencial de primeira ordem com valor inicial:

$$\frac{dy(t)}{dt} = f(y(t),t) = y(t)(1-y(t)) \quad (21.3)$$
$$y(0) = 1/2 \quad (21.4)$$

A solução deste problema é

$$y(t) = \frac{e^t}{1+e^t}$$

Para este problema

$$y(t+h) = y(t) + hf(t) = y(t) + h(y(t)(1-y(t))) = (1+h)y(t) - h(y(t))^2$$

Resolva este problema com

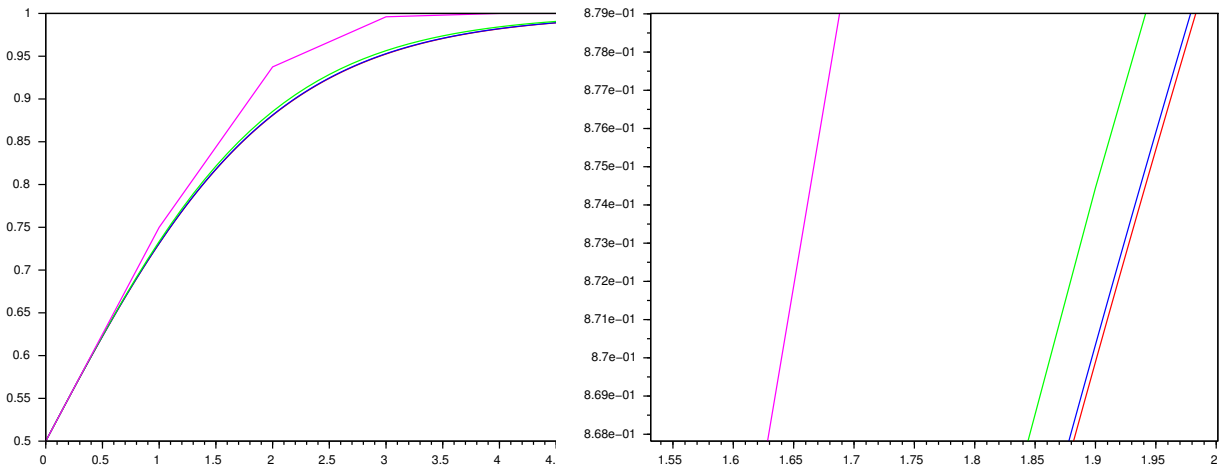
a) $h = 0{,}1$

b) $h = 0{,}01$

c) $h = 1$

Solução:

Para o caso "a" são usados 50 pontos, para o caso "b" são usados 500 pontos e para o caso "c" são usados 5 pontos.



(a) Soluções (b) Zoom

Figura 21.1: Soluções do problema: em vermelho solução analítica, em azul solução numérica com $h = 0{,}01$,em verde solução numérica com $h = 0{,}1$ e em magenta solução numérica com $h = 1$,

21.3 Método de Euler-melhorado

O método de Euler utiliza uma aproximação para a derivada da função cujo erro é proporcional ao tamanho do passo h. Quando h é bastante pequeno a solução se aproxima da solução analítica, mas na medida em que h cresce, a solução se distancia da solução analítica.

Uma maneira de melhorar o método de Euler consiste em utilizar outra aproximação para a derivada da função. No exemplo acima, se fosse utilizado um valor um pouco menor no lugar de $f(t)$, teríamos uma solução melhor.

O método de Euler-melhorado consistem em dois passos. Inicialmente, utiliza-se o método de Euler com apenas meio passo $h/2$.

$$y_{meio} = y(t + h/2) = y(t) + \frac{h}{2} f(y(t), t)$$

Depois, utiliza-se o método de Euler com passo inteiro, onde a derivada da função é calculada no ponto meio passo a frente.

$$\begin{aligned} y(t+h) &= y(t) + hf(y(t+h/2), t+h/2) && (21.5) \\ &= y(t) + hf(y_{meio}, t+h/2) && (21.6) \end{aligned}$$

21.3.1 Exemplo 3

Considere novamente a seguinte equação diferencial de primeira ordem com valor inicial:

$$\begin{aligned} \frac{dy(t)}{dt} &= f(y(t), t) = y(t)(1 - y(t)) && (21.7) \\ y(0) &= 1/2 && (21.8) \end{aligned}$$

Resolva este problema com $h = 1$ utilizando 5 pontos do Método de Euler-melhorado

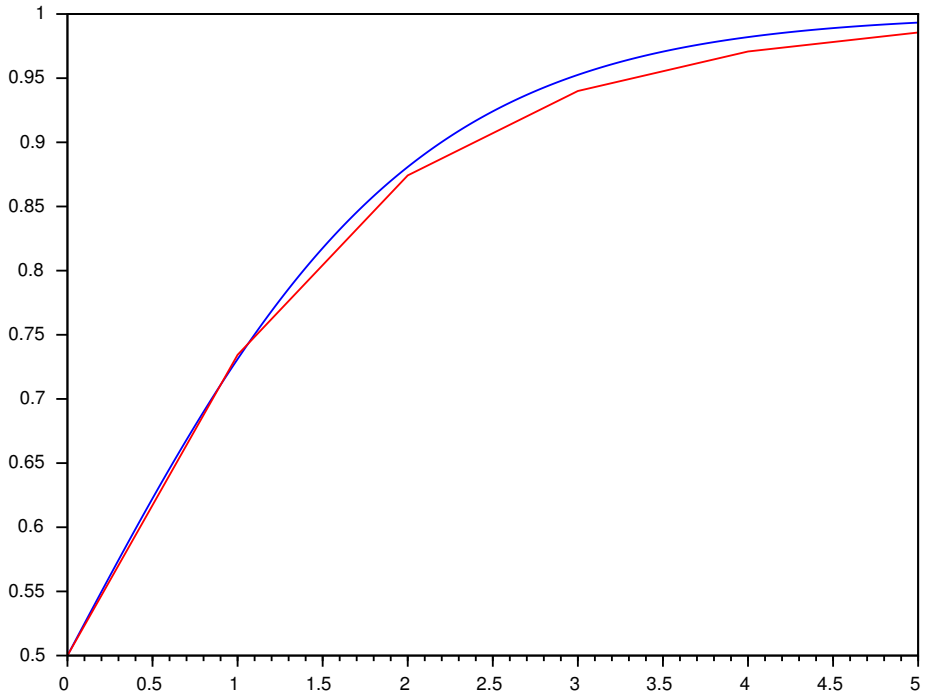


Figura 21.2: Soluções do problema: em zul solução analítica e em vermelho solução com método Euler-melhorado

21.4 Exercícios

Exercise 43 Resolva a seguinte equação diferencial, utilizando os métodos de Euler e Euler-

melhorado. Escolha adequadamente o tamanho do passo h.

$$\frac{dy(t)}{dt} = \frac{t - y(t)}{t} \tag{21.9}$$

$$y(2) = 3 \tag{21.10}$$

Aula 22

Runge-Kutta e Adams-Bashforth

22.1 Método de Runge-Kutta

Considere a seguinte equação diferencial de primeira ordem com valor inicial:

$$\frac{dy(t)}{dt} = f(y(t), t) \tag{22.1}$$
$$y(t_0) = y_0 \tag{22.2}$$

Gostaríamos de obter uma aproximação para a solução $y(t)$ deste problema, a qual pode ser descrita pela série de Taylor. A série de Taylor associada a uma função f infinitamente diferenciável definida em um intervalo aberto $(a - r, a + r)$ é a série de potências dada por

$$y(t + h) = \sum_{n=0}^{\infty} \frac{y^{(n)}(t)}{n!}(h)^n,$$

onde, $n!$ é o fatorial de n e $f^{(n)}(t)$ denota a n-ésima derivada de f no ponto t.

Uma aproximação para a função pode ser obtida truncando a série no n-ésimo termo.

$$y(t + h) = y(t) + y'(t)(h) + \cdots + \frac{y^{(n)}(t)}{n!}(h)^n + R(t)$$

onde $R(t)$ representa o erro de truncamento.

A idéia principal do **Método de Runge-Kutta** consistem em aproximar a solução da equação diferencial por

$$y(t + h) = y(t) + h\sum_{i=1}^{m} b_i k_i \tag{22.3}$$
$$k_1 = f(y(t), t) \tag{22.4}$$
$$k_i = f\left(y(t) + h\sum_{j=1}^{i-1} a_{ij}k_j, t + c_i h\right), \quad i = 2, \ldots, m. \tag{22.5}$$

Para tanto, precisamos calcular as constantes b_i, c_i e a_{ij} que melhor aproximam a solução ao resultado analítico.

22.2 Runge-Kutta de Segunda Ordem

O Método Runge-Kutta de segunda ordem ($m = 2$) é definido pelas seguintes relações:

$$b_1 + b_2 = 1 \tag{22.6}$$
$$b_2 c_2 = 1/2 \tag{22.7}$$
$$b_2 a_{21} = 1/2 \tag{22.8}$$

22.2.1 Método de Euler-Melhorado

$$b_1 = 0 \tag{22.9}$$
$$b_2 = 1 \tag{22.10}$$
$$c_2 = 1/2 \tag{22.11}$$
$$a_{21} = 1/2 \tag{22.12}$$

Portanto

$$\begin{aligned} k_1 &= f(y(t), t) && (22.13)\\ k_2 &= f\left(y(t) + \frac{1}{2}hk_1, t + \frac{1}{2}h\right) && (22.14)\\ y(t+h) &= y(t) + hk_2 && (22.15) \end{aligned}$$

22.2.2 Método Predição-Correção

$$b_1 = 1/2 \tag{22.16}$$
$$b_2 = 1/2 \tag{22.17}$$
$$c_2 = 1 \tag{22.18}$$
$$a_{21} = 1 \tag{22.19}$$

Portanto

$$\begin{aligned} k_1 &= f(y(t), t) && (22.20)\\ k_2 &= f\left(y(t) + hk_1, t + h\right) && (22.21)\\ y(t+h) &= y(t) + h\frac{1}{2}k_1 + h\frac{1}{2}k_2 && (22.22) \end{aligned}$$

22.3 Runge-Kutta de Quarta Ordem

$$\begin{aligned} k_1 &= f(y(t), t) && (22.23)\\ k_2 &= f\left(y(t) + hk_1/2, t + h/2\right) && (22.24)\\ k_3 &= f\left(y(t) + hk_2/2, t + h/2\right) && (22.25)\\ k_4 &= f\left(y(t) + hk_3, t + h\right) && (22.26)\\ y(t+h) &= y(t) + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k4\right) && (22.27) \end{aligned}$$

22.4 Método de Adams-Bashforth

22.4.1 Adams-Bashforth de segunda ordem

$$y(t+h) = y(t) + \frac{h}{2}\left(3f(y(t), t) - f(y(t-h), t-h)\right)$$

22.4.2 Adams-Bashforth de terceira ordem

$$y(t+h) = y(t) + \frac{h}{12}\left(23f(y(t), t) - 16f(y(t-h), t-h) + 5f(y(t-2h), t-2h)\right)$$

22.4.3 Adams-Bashforth de quarta ordem

$$y(t+h) = y(t) + \frac{h}{24}\left(55f(y(t), t) - 59f(y(t-h), t-h) + 37f(y(t-2h), t-2h) - 9f(y(t-3h), t-3h)\right)$$

22.5 Erros

Método	Erro
Euler	$\mathcal{O}(h)$
Euler-Melhorado	$\mathcal{O}(h^2)$
Predição-Correção	$\mathcal{O}(h^2)$
Runge-Kutta 4	$\mathcal{O}(h^4)$
Adams-Bashforth 2	$\mathcal{O}(h^2)$
Adams-Bashforth 3	$\mathcal{O}(h^3)$
Adams-Bashforth 4	$\mathcal{O}(h^4)$

Aula 23

Equações diferenciais de segunda ordem

23.1 Equações diferenciais de segunda ordem

Um objeto inicialmente parado cai de uma altura de 100 metros. Qual a altura após 1 segundo? Considere o atrito com o ar.

$$\frac{dv(t)}{dt} = -9{,}8 + 0{,}1(v(t))^2$$

Primeiramente é necessário encontrar uma equação que relacione a velocidade com a altura. Das aulas de física sabemos que

$$\frac{dh(t)}{dt} = v(t)$$

e derivando os dois lados da equação temos

$$\frac{d^2h(t)}{dt^2} = \frac{dv(t)}{dt}.$$

Podemos agora substituir estas equações da altura na equação diferencial inicial:

$$\frac{d^2h(t)}{dt^2} = -9{,}8 + 0{,}1\left(\frac{dh(t)}{dt}\right)^2$$

Temos uma equação diferencia de segunda ordem. Para resolve-la precisamos de 2 condições iniciais. Sabemos que a altura inicial é 100 metros e que a velocidade inicial é 0 metros por segundo. Logo,

$$\frac{d^2h(t)}{dt^2} - 0{,}1\left(\frac{dh(t)}{dt}\right)^2 + 9{,}8 = 0 \tag{23.1}$$

$$h(0) = 100 \tag{23.2}$$

$$h'(0) = 0 \tag{23.3}$$

23.2 Sistema de equações de primeira ordem

Todas as técnicas aprendidas até agora podem ser utilizadas em equações diferenciais de primeira ordem. Portanto, vamos transformar a equação diferencial de segunda ordem em duas equações diferencias de primeira ordem para poder utilizar todos os algoritmos desenvolvidos até agora.

Vamos utilizar uma variável auxiliar $z(t)$ para criar a nossa primeira equação diferencial

$$h'(t) = z(t)$$

e derivando dos dois lados da equação

$$h''(t) = z'(t).$$

Substituindo a variável auxiliar na equação original temos a segunda equação diferencial

$$z'(t) - 0{,}1\,(z(t))^2 + 9{,}8 = 0. \tag{23.4}$$

A segunda equação diferencial é

$$h'(t) = z(t)$$

Juntando as duas e as condições de contorno temos que

$$h'(t) = z(t) \tag{23.5}$$
$$z'(t) - 0{,}1\,(z(t))^2 + 9{,}8 = 0 \tag{23.6}$$
$$h(0) = 100 \tag{23.7}$$
$$z(0) = 0 \tag{23.8}$$

Temos agora duas equações diferenciais de primeira ordem e portanto podemos usar agora os algoritmos para resolver estas equações.

23.3 Sistema vetorial

Para simplificar a solução das equações diferencias podemos fazer mais uma mudança de variáveis.

$$y(t) = \begin{bmatrix} h(t) \\ z(t) \end{bmatrix}$$

e reescrever a equação diferencial como

$$y'(t) = \begin{bmatrix} h'(t) \\ z'(t) \end{bmatrix} = \begin{bmatrix} -z(t) \\ +0{,}1\,(z(t))^2 - 9{,}8 \end{bmatrix} = \begin{bmatrix} -y_2(t) \\ +0{,}1\,(y_2(t))^2 - 9{,}8 \end{bmatrix} \tag{23.9}$$
$$y(0) = \begin{bmatrix} h(0) \\ z(0) \end{bmatrix} = \begin{bmatrix} 100 \\ 0 \end{bmatrix} \tag{23.10}$$

Simplificando

$$y'(t) = F(y,t) = \begin{bmatrix} -y_2(t) \\ +0{,}1\,(y_2(t))^2 - 9{,}8 \end{bmatrix} \tag{23.11}$$
$$y(0) = \begin{bmatrix} 100 \\ 0 \end{bmatrix} \tag{23.12}$$

Portanto temos 1 equação diferencial vetorial de primeira ordem e podemos utilizar todos os algoritmos aprendidos até agora.

23.4 Conjunto de equações diferenciais

Observe que em muitos problemas nós já temos no início equações diferencias de primeira ordem. Nós sabíamos que

$$\frac{dv(t)}{dt} = -9{,}8 + 0{,}1(v(t))^2$$

e que $v(0) = 0$.

Além disso nós sabíamos que

$$\frac{dh(t)}{dt} = v(t)$$

e que $h(0) = 100$.

Nesse caso basta escolher $y(t)$ e reescrever o problema na forma matricial.

$$y(t) = \begin{bmatrix} v(t) \\ h(t) \end{bmatrix}$$

$$y'(t) = F(y,t) = \begin{bmatrix} -9{,}8 + 0{,}1\,(y_1(t))^2 \\ -y_1(t) \end{bmatrix} \tag{23.13}$$

$$y(0) = \begin{bmatrix} 0 \\ 100 \end{bmatrix} \tag{23.14}$$

Aula 24

Equação do calor em regime estacionário

24.1 Equação do calor em regime estacionário

Para desenvolver a equação do calor vamos utilizar algumas leis da física e vamos fazer algumas hipóteses sobre o modelo.

Vamos considerar que a temperatura da barra está em regime estacionário, e portanto a variação da energia é zero.

$$\frac{dE}{dt} = 0 = \sum q_i - \sum q_o$$

Onde q_i são os fluxos de energia que entram na barra e q_o são os fluxos de energia que saem da barra.

Vamos considerar que em uma seção da barra ocorrem 4 fluxos de energia:

- Condução no lado esquerdo da seção;
- Condução no lado direito da seção;
- Radiação;
- Transferência de energia através de uma fonte de calor P.

Segundo a Lei de Fourier, o fluxo de energia por **condução** se dá segundo:

$$q = -kA\frac{dT(x)}{dx}$$

onde A é a área da seção transversal e k é resistência térmica. Esta lei significa que o calor flui do meio com maior temperatura para o meio com menor temperatura.

Uma seção da barra transfere energia por **radiação** segundo:

$$q = \epsilon\sigma((T(x))^4 - T_a^4)$$

onde T_a é a temperatura do meio ambiente, σ é constante de Stefan-Boltzmann e ϵ é a *emissividade* que depende do tamanho da superfície da seção. Se $T \gg T_a$ a expressão pode ser simplificada como $q = \epsilon\sigma T(x)^4$.

Se considerarmos uma seção da barra com tamanho Δx temos que

$$\frac{dE}{dt} = -kA\frac{dT(x)}{dx} + kA\frac{dT(x+\Delta x)}{dx} - \epsilon\sigma(T(x))^4 + P = 0$$

Dividindo a equação por Δx:

$$\frac{kA\frac{dT(x+\Delta x)}{dx} - kA\frac{dT(x)}{dx} +}{\Delta x} - \frac{\epsilon}{\Delta x}\sigma(T(x))^4 + \frac{P}{\Delta x} = 0$$

E fazendo o limite de Δx tendendo a zero temos que

$$k_1 \frac{d^2T(x)}{dx^2} - k_2(T(x))^4 + dP = 0$$

Onde k_1 e k_2 são constantes que dependem do material que a barra é feita e dP é densidade de potência por unidade de comprimento.

Vamos considerar 3 tipos de condições de contorno:

- Temperatura conhecida

  Se em uma extremidade x_0 a temperatura T_0 é conhecida devemos utilizar como condição de contorno

  $$T(x_0) = T_0$$

- Extremidade isolada

  Se uma das extremidades está isolada então não ocorre transferência de energia por condução o que significa que $q = 0$ e portanto

  $$\left.\frac{dT}{dx}\right|_{x=x_0} = 0$$

- Fonte de calor na extremidade

  Caso exista uma fonte de calor Q_0 na extremidade é conhecido o fluxo de energia que entra na barra por condução, logo $-kA\frac{dT(x)}{dx} = Q_0$ e portanto

  $$\left.\frac{dT(x)}{dx}\right|_{x=x_0} = -\frac{Q_0}{kA} = -\frac{Q_0}{k_1}$$

24.2 Solução por diferenças finitas

24.2.1 Caso 1

Vamos inicialmente considerar que não existe fonte de calor fornecendo energia para a barra ($dP = 0$) e que não ocorre perda de calor por radiação ($k_2 = 0$). Neste caso a equação diferencial é simplificada como:

$$k_1 \frac{d^2T(x)}{dx^2} = 0 \quad \Leftrightarrow \quad \frac{d^2T(x)}{dx^2} = 0$$

Vamos considerar que as temperatura nas extremidades são conhecidas, logo:

$$T(x_0) = T_0 \quad \text{e} \quad T(x_1) = T_1.$$

Como exemplo, vamos considerar que o comprimento da barra é 2 metros, e que a temperatura das extremidades é 100 e 120 graus. Neste caso, o problema se torna

$$\begin{aligned} &\frac{d^2T(x)}{dx^2} = 0 \\ &T(0) = 100 \\ &T(2) = 120 \end{aligned}$$

Vamos dividir a barra em 10 elementos, de maneira que gostaríamos de saber a temperatura em $T(0)$, $T(0{,}2)$, $T(0{,}4)$, $\cdots$, $T(1.8)$, $T(2)$. Para calcular a temperatura nos 11 pontos, precisamos de 11 equações. Duas delas são:

$$T(0) = 100 \quad \text{e} \quad T(2) = 120,$$

precisamos ainda de mais 9 equações. Aproximando a segunda derivada na equação diferencial acima temos que:

$$\frac{d^2T(x)}{dx^2} \approx \frac{T(x-h) - 2T(x) + T(x+h)}{h^2} = 0 \quad \Leftrightarrow T(x-h) - 2T(x) + T(x+h) = 0.$$

Podemos escrever a equação acima para $x = 0{,}2$, $x = 0{,}4$, $x = 0{,}6$, $\cdots$, $x = 1{,}8$ e assim obter mais 9 equações:

$$\begin{gathered} T(0) - 2T(0{,}2) + T(0{,}4) = 0 \\ T(0{,}2) - 2T(0{,}4) + T(0{,}6) = 0 \\ T(0{,}4) - 2T(0{,}6) + T(0{,}8) = 0 \\ \vdots \\ T(1{,}6) - 2T(1{,}8) + T(2) = 0 \end{gathered}$$

Agrupando todas as equações temos que:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & & 0 \\ 0 & 1 & -2 & 1 & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} T(0) \\ T(0{,}2) \\ T(0{,}4) \\ \vdots \\ T(1{,}8) \\ T(2) \end{pmatrix} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 120 \end{pmatrix}$$

cuja solução é

$$\begin{pmatrix} T(0) \\ T(0{,}2) \\ T(0{,}4) \\ \vdots \\ T(1{,}8) \\ T(2) \end{pmatrix} = \begin{pmatrix} 100 \\ 102 \\ 104 \\ \vdots \\ 118 \\ 120 \end{pmatrix}$$

24.2.2 Caso 2

Vamos agora considerar um caso mais completo, onde a temperatura na extremidade esquerda é conhecida $T(0) = 100$ e na extremidade direita existe uma fonte de calor $Q_0 = 100$. Além disso, vamos considerar que a barra perde calor por radiação e que k_1=1 e $k_2 = 10^{-7}$. Neste caso o equação diferencial é:

$$1\frac{d^2T(x)}{dx^2} - 10^{-7}(T(x))^4 = 0$$

e as condições de contorno são

$$T(0) = 100 \quad \text{e} \quad \left.\frac{dT}{dx}\right|_{x=2} = 100$$

Vamos novamente dividir a barra em 10 elementos, de maneira que gostaríamos de saber a temperatura em $T(0)$, $T(0{,}2)$, $T(0{,}4)$, $\cdots$, $T(1.8)$, $T(2)$. Para calcular a temperatura nos 11 pontos, precisamos de 11 equações. A primeira equação vem da condição de contorno na esquerda

$$T(0) - 100 = 0.$$

A segunda equação vem da condição de contorno na direita. Vamos aproximar a derivada na direita usando a técnica *diferenças regressivas*. Logo

$$\frac{T(2) - T(1{,}8)}{0{,}2} = 100 \Leftrightarrow T(2) - T(1{,}8) - 20 = 0$$

Aproximando a segunda derivada na equação diferencial (usando a técnica das *diferenças centrais*) temos que:

$$\frac{d^2T(x)}{dx^2} \approx \frac{T(x-h)-2T(x)+T(x+h)}{h^2} - k_2T(x)^4 = 0 \quad \Leftrightarrow T(x-h)-2T(x)+T(x+h)-h^2k_2T(x)^4 = 0.$$

Podemos escrever a equação acima para $x = 0{,}2$, $x = 0{,}4$, $x = 0{,}6$, $\cdots$, $x = 1{,}8$ e assim obter mais 9 equações:

$$\begin{gathered}
T(0) - 2T(0{,}2) + T(0{,}4) - 0{,}2^2k_2T(0{,}2)^4 = 0 \\
T(0{,}2) - 2T(0{,}4) + T(0{,}6) - 0{,}2^2k_2T(0{,}4)^4 = 0 \\
T(0{,}4) - 2T(0{,}6) + T(0{,}8) - 0{,}2^2k_2T(0{,}6)^4 = 0 \\
\vdots \\
T(1{,}6) - 2T(1{,}8) + T(2) - 0{,}2^2k_2T(1{,}8)^4 = 0
\end{gathered}$$

Agrupando todas as equações temos:

$$F(T) = 0 \Leftrightarrow \begin{cases}
T(0) - 100 = 0 \\
T(0) - 2T(0{,}2) + T(0{,}4) - 0{,}2^2k_2T(0{,}2)^4 = 0 \\
T(0{,}2) - 2T(0{,}4) + T(0{,}6) - 0{,}2^2k_2T(0{,}4)^4 = 0 \\
T(0{,}4) - 2T(0{,}6) + T(0{,}8) - 0{,}2^2k_2T(0{,}6)^4 = 0 \\
\vdots \\
T(1{,}6) - 2T(1{,}8) + T(2) - 0{,}2^2k_2T(1{,}8)^4 = 0 \\
T(2) - T(1{,}8) - 20 = 0
\end{cases}$$

que é um conjunto de 11 equações não lineares $F(T) = 0$. Para resolver estas equações podemos utilizar o Método de Newton. Para tanto precisamos calcular o Jacobiano de $F(T)$.

$$J(T) = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
1 & -2 - 0{,}2^2 \cdot k_2 4T(0{,}2)^3 & 1 & & 0 \\
0 & 1 & -2 - 0{,}2^2 \cdot k_2 4T(0{,}4)^3 & 1 & \\
\vdots & & \ddots & \ddots & \\
0 & \cdots & 1 & -2 - 0{,}2^2 \cdot k_2 4T(1{,}8)^3 & 1 \\
0 & \cdots & 0 & -1 & 1
\end{pmatrix}$$

Utilizando a condição inicial

$$T = \begin{pmatrix} 100 & 100 & 100 & 100 & 100 & 100 & 100 & 100 & 100 & 100 & 100 \end{pmatrix}^T$$

após 10 iterações do algoritmo de Newton chegamos ao resultado

$$T = \begin{pmatrix} 100. & 105.1371713316 & 110.7630908336 & 116.9910710036 & 123.9683772711 & 131.890404729 \\ 141.022787261 & 151.7372087285 & 164.5720817492 & 180.3411205134 & 200.3411205134 \end{pmatrix}^T$$

Parte VII

Apêndice

Apêndice A

Teoremas

A.1 Teorema de Rolle

Seja $f : [a, b] \to \mathbb{R}$ uma função contínua no intervalo fechado $[a, b]$, e diferenciável no intervalo (a, b), onde $a < b$. Se $f(a) = f(b)$ então existe um valor c no intervalo (a, b) tal que:

$$f'(c) = 0$$

Prova: Por contradição.

Se não existe ponto c tal que $f'(x) = 0$, então OU $f'(x) > 0$ OU $f'(x) < 0$.

Se $f'(x) > 0$ a função é crescente e portanto $f(b) > f(a)$. Se $f'(x) < 0$ a função é decrescente e portanto $f(a) > f(b)$. Portanto deve existir c tal que $f'(x) = 0$.

A.2 Teorema do Valor Médio

Seja $f : [a, b] \to \mathbb{R}$ uma função contínua no intervalo fechado $[a, b]$, e diferenciável no intervalo (a, b), onde $a < b$. Então existe um valor c no intervalo (a, b) tal que:

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

Prova:

Defina

$$g(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a)$$

Observe que $g(b) = 0$ e $g(a) = 0$, portanto pelo Teorema de Rolle existe $c \in (a, b)$ tal que $g'(c) = 0$

Calculando $g'(c)$ e igualando a zero

$$g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0$$

A.3 Teorema de Cauchy

Seja $f : [a, b] \to \mathbb{R}$ e $g : [a, b] \to \mathbb{R}$ funções contínuas no intervalo fechado $[a, b]$, e diferenciáveis no intervalo (a, b), onde $a < b$. Então existe um valor c no intervalo (a, b) tal que:

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

Prova:

Defina $h = f - rg$, onde

$$r = \frac{f(b) - f(a)}{g(b) - g(a)}$$

logo $h(a) = h(b)$.

Derivando h temos que $h' = f' - rg'$. Utilizando o Teorema do valor médio em h, temos que

$$\exists c \in (a,b) : h'(c) = \frac{f(b) - f(a)}{b - a}$$

Portanto

$$h'(c) = 0 = f'(c) - rg(c)$$

Logo

$$\frac{f'(c)}{g'(c)} = r = \frac{f(b) - f(a)}{g(b) - g(a)}$$

que conclui a prova.

A.4 Teorema de Taylor

Pela Série de Taylor

$$f(x+h) = \sum_{k=0}^{\infty} \frac{f^k(x)h^k}{k!}$$

Que pode ser truncada em n

$$f(x+h) = \sum_{k=0}^{n} \frac{f^k(x)h^k}{k!} + R$$

Onde R é a continuação da série:

$$R = f(x+h) - \sum_{k=0}^{n} \frac{f^k(x)h^k}{k!}$$

Utilizando a Série de Taylor novamente

$$f(x+h) = \sum_{k=0}^{\infty} \frac{f^k(t)(x+h-t)^k}{k!} = \sum_{k=0}^{n} \frac{f^k(t)(x+h-t)^k}{k!} + S(t)$$

onde

$$S(t) = f(x+h) - \sum_{k=0}^{n} \frac{f^k(t)(x+h-t)^k}{k!}$$

Observe que $S(x+h) = 0$ e que $S(x) = R$.

Calculando a derivada de S com relação a t (E SIMPLIFICANDO O SOMATÓRIO) temos que

$$\frac{dS}{dt} = -\frac{f^{n+1}(t)(x+h-t)^n}{n!}$$

Vamos agora definir a função $g(t) = (x+h-t)^{n+1}$. Sabemos que

$$\begin{aligned} g(x+h) &= 0 \\ g(x) &= h^{n+1} \\ \frac{dg}{dt} &= -(n+1)(x+h-t)^n \end{aligned}$$

Utilizando o Teorema de Cauchy, sabemos que existe $c \in (x, x+h)$ tal que

$$\frac{S'(c)}{g'(c)} = \frac{S(x+h) - S(x)}{g(x+h) - g(x)}$$

Portanto,

$$\frac{\left(-\frac{f^{(n+1)}(c)(x+h-c)^n}{n!}\right)}{(-(n-1)(x+h-c)^n)} = \frac{0-R}{0-h^{n+1}}$$

$$\frac{\left(-\frac{f^{(n+1)}(c)}{n!}\right)}{(-(n+1))} = \frac{0-R}{0-h^{n+1}}$$

Logo

$$R = \frac{f^{(n+1)}(c)h^{n+1}}{(n+1)!}$$