Cálculo Numérico

Um Livro Colaborativo

6 de julho de 2016

Autores

Lista de autores[1]:

Esequia Sauter - UFRGS

Fabio Souto de Azevedo - UFRGS

Pedro Henrique de Almeida Konzen - UFRGS

[1] em ordem alfabética

Licença

Nota dos autores

Este livro vem sendo construído de forma colaborativa desde 2011. Nosso intuito é de melhorá-lo, expandi-lo e adaptá-lo às necessidades de um curso semestral de cálculo numérico em nível de graduação.

Caso queira colaborar, entre em contato conosco pelo endereço de e-mail:

`livro_colaborativo@googlegroups.com`

Apresentação

Este livro busca abordar os tópicos de um curso de introdução ao cálculo numérico moderno oferecido a estudantes de matemática, física, engenharias e outros. A ênfase é colocada na formulação de resolução de problemas, implementação em computador e interpretação de resultados. Pressupõe-se que o estudante domine conhecimentos e habilidades típicas desenvolvidas em cursos de graduação de cálculo, álgebra linear e equações diferenciais. Conhecimentos prévios em linguagem de computadores é fortemente recomendável, embora apenas técnicas elementares de programação sejam realmente necessárias.

Sumário

Capítulo 1

Introdução

Cálculo numérico é uma disciplina que compreende o estudo de métodos para a computação eficiente da solução de problemas matemáticos. Aliado ao avanço tecnológico dos computadores, o desenvolvimento de métodos numéricos tornou a simulação computacional de modelos matemáticos uma prática cotidiana nas mais diversas áreas científicas e tecnológicas. As então chamadas simulações numéricas são constituídas de um arranjo de vários esquemas numéricos dedicados a resolver problemas específicos como, por exemplo: resolver equações algébricas, resolver sistemas lineares, interpolar e ajustar pontos, calcular derivadas e integrais, resolver equações diferenciais ordinárias, etc.. Neste livro, abordamos o desenvolvimento, a implementação, utilização e aspectos teóricos de métodos numéricos para a resolução desses problemas.

Os problemas que discutiremos não formam apenas um conjunto de métodos fundamentais, mas são, também, problemas de interesse na engenharia e na matemática aplicada. Estes podem se mostrar intratáveis se dispomos apenas de meios puramente analíticos, como aqueles estudados nos cursos de cálculo e álgebra linear. Por exemplo, o teorema de Abel-Ruffini nos garante que não existe uma fórmula algébrica, isto é, envolvendo apenas operações aritméticas e radicais, para calcular as raízes de uma equação polinomial de qualquer grau, mas apenas casos particulares:

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

- Existe expressão para equações de quarto grau;
- Casos simplificados de equações de grau maior que 4 onde alguns coeficientes são nulos também podem ser resolvidos.

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

$$\cos(x) = x \qquad \text{e} \qquad xe^x = 10$$

Para resolver o problema de valor inicial

$$\begin{cases} y' + xy = x, \\ y(0) = 2, \end{cases}$$

podemos usar o método de fator integrante e obtemos $y = 1 + e^{-x^2/2}$. Já o cálculo da solução exata para o problema

$$\begin{cases} y' + xy = e^{-y}, \\ y(0) = 2, \end{cases}$$

não é possível.

Da mesma forma, resolvemos a integral

$$\int_1^2 xe^{-x^2} dx$$

pelo método da substituição e obtemos $\frac{1}{2}(e^{-1} - e^{-2})$. Porém a integral

$$\int_1^2 e^{-x^2} dx$$

não pode ser resolvida analiticamente.

A maioria das modelagem de fenômenos reais chegam em problemas matemáticos onde a solução analítica é difícil (ou impossível) de ser encontrada, mesmo quando provamos que ela existe. Nesse curso propomos calcular aproximações numéricas para esses problemas, que apesar de, em geral, serem diferentes da solução exata, mostraremos que elas podem ser bem próximas.

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

Capítulo 2

Aritmética de Máquina

2.1 Sistema de Numeração e Mudança de Base

Usualmente, utilizamos o sistema de numeração decimal para representar números. Esse é um sistema de numeração posicional onde a posição do dígito indica a potência de 10 que o dígito está representando.

Exemplo 1. *O número 293 decomposto em centenas, dezenas e unidades:*

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

Assim, vemos que as centenas, dezenas e unidades são potências de 10.

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

Definição 1 (Sistema de numeração de base b)**.** *Dado um número natural $b > 1$ e a coleção de símbolos $\{\text{“,”}, -, 0, 1, 2, \ldots, b-1\}$[1], a sequência de dígitos:*

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

representa o número positivo

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

Observação 1 ($b \geq 10$)**.** *Para sistemas de numeração com base $b \geq 10$ é usual utilizar as seguintes notações:*

[1] Para sistemas de numeração com base $b > 10$, veja a Observação 1

- *No sistema de numeração decimal, i.e.* $b = 10$*, representamos o número:*

$$\pm(d_n d_{n-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots)_{10}$$

simplesmente por:

$$\pm d_n d_{n-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots$$

Ou seja, não usamos parênteses, nem o subíndice indicando a base.

- *Em sistemas de numeração com base* $b > 10$*, usamos as letras* A*,* B*,* C*, etc., para denotar os símbolos:* $A = 10$*,* $B = 11$*,* $C = 12$*, etc..*

Exemplo 2 (Sistema binário)**.** *O sistema de numeração em base dois é chamado de binário e os algarismos binários são conhecidos como bits, do inglês* **binary digits***. Um bit pode assumir apenas dois valores distintos:* 0 *ou* 1*. Por exemplo:*

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

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

Exemplo 3 (Sistema quaternário)**.** *No sistema quaternário a base* b *é igual a* 4*. Por exemplo:*

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

Exemplo 4 (Sistema octal)**.** *No sistema quaternário a base é* $b = 8$*. Por exemplo:*

$$\begin{aligned}(1357,24)_8 &= 1\cdot 8^3 + 3\cdot 8^2 + 5\cdot 8^1 + 7\cdot 8^0 + 2\cdot 8^{-1} + 4\cdot 8^{-2} \\ &= 512 + 192 + 40 + 7 + 0,25 + 0,0625 = (751,3125)_{10}\end{aligned}$$

Exemplo 5 (Sistema hexadecimal)**.** *O sistema de numeração cuja a base é* $b = 16$ *é chamado de sistema hexadecimal. O conjunto de símbolos necessários é* $S = \{\text{“, ”}, -, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$*. O número* $(E2AC)_{16}$ *no sistema decimal é igual a:*

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

Exercício 1. *Escreva os números abaixo na base decimal*

a) $(25, 13)_8$

b) $(101, 1)_2$

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

d) $(11, 2)_3$

A partir da Definição 1 acabamos de mostrar vários exemplos de conversão de números de uma sistema de numeração de base b para o sistema decimal. Agora, vamos estudar como fazer o processo inverso. Isto é, dado um número decimal queremos escrevê-lo em uma outra base b. Para tanto, consideramos um número decimal X_{10} representado na base b:

$$\begin{aligned} X_{10} &= (d_n d_{n-1} \cdots d_0, d_{-1} \cdots)_b \\ &= d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_1 \cdot b^1 + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots \end{aligned}$$

Separando as partes inteira e parte fracionária de X, i.e. $X = X^i + X^f$, temos:

$$X^i = d_n \cdot b^n + \cdots + d_{n-1} b^{n-1} \cdot + d_1 \cdot b^1 + d_0 \cdot b^0 \qquad \text{e} \qquad X^f = \frac{d_{-1}}{b^1} + \frac{d_{-2}}{b^2} + \cdots$$

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

Primeiramente, vejamos como tratar a parte inteira X^i. Calculando sua divisão por b, temos:

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

Observe que d_0 é o resto da divisão de X^i por b, pois $d_1 + d_2 b^1 \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ é inteiro e $\frac{d_0}{b}$ é uma fração (lembramos que $d_0 < b$). Da mesma forma, o resto da divisão de $d_1 + d_2 b^1 \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ por b é d_1. Repetimos o processo até encontrar os símbolos $d_0, d_1, d_2, \ldots$.

Exemplo 6 (Conversão da parte inteira)**.** *Vamos escrever o número 125 na base* 6*. Para encontrar* d_0*, dividimos* 125 *por 6:*

$$\begin{array}{r|l} 125 & 6 \\ \underline{12} & 20 \\ 05 & \\ \underline{00} & \\ 5 & \end{array}$$

e encontramos $d_0 = 5$. *Dividindo o quociente por* 6 *para encontrar* d_1*:*

$$\begin{array}{r|l} 20 & 6 \\ \underline{18} & 3 \\ 2 & \end{array}$$

e obtemos $d_1 = 2$. *Observe que o quociente agora é menor que* 6*, ou seja, uma sucessiva divisão por* 6 *teria resto igual ao próprio quociente. Assim, concluímos que:*

$$125 = (325)_6$$

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

```
-->q = 125, d0 = modulo(q,6)
-->q = int(q/6), d1 = modulo(q,6)
-->q = int(q/6), d2 = modulo(q,6)
```

Verifique!

Agora, para convertermos a parte fracionária X^f na base b, i.e. para encontrar os símbolos d_{-1}, d_{-2}, etc, multiplicamos a parte fracionária de X por b:

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

Observe que a parte inteira desse produto é d_{-1} e $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ é a parte fracionária. Quando multiplicamos $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ por b novamente, encontramos d_{-2}. Repetimos o processo até encontrar todos os símbolos.

Exemplo 7 (Conversão da parte fracionária)**.** *Escrever o número* $125,58\overline{3}$ *na base* 6*. Do exemplo anterior temos que* $125 = (325)_6$*. Assim, nos resta converter a parte fracionária, Multiplicando-a por* 6*:*

$$\begin{array}{r} 0,58\overline{3} \\ \times\ 6 \\ \hline 3,49\overline{9} \end{array}$$

e obtemos $d_{-1} = 3$. *Agora, multiplicamos* $0,49\overline{9} = 0,5$ *por* 6*:*

$$\begin{array}{r} 0,5 \\ \times\ 6 \\ \hline 3,0 \end{array}$$

e obtemos $d_{-2} = 3$. *Portanto:*

$$125,58\overline{3} = (325,33)_6$$

As contas feitas aqui, também podem ser feitas no Scilab. Você sabe como?

Exercício 2. *Escreva cada número decimal na base b*

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

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

Uma maneira de converter um número dado numa base g para uma base b é fazer em duas partes: primeiro converter o número dado na base g para base decimal e depois converter para a base b.

Exercício 3. *Escreva cada número dado para a base b.*

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

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

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

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

2.2 Aritmética de Máquina

Os computadores, em geral, usam uma base binária para representar os números, onde as posições, chamadas de bits, assume as condições "verdadeiro" ou "falso", ou seja, 0 ou 1. Cada computador tem um número de bits fixo e, portanto, representa uma quantidade finita de números. Os demais números são tomados por proximidade àqueles conhecidos, gerando erros de arredondamento. Por exemplo, em aritmética de computador, o número 2 tem representação exata, logo $2^2 = 4$, mas $\sqrt{3}$ não tem representação finita, logo $(\sqrt{3})^2 \neq 3$. Veja isso no Scilab:

```
-->2^2 == 4
 ans  =
  T
-->sqrt(3)^2 == 3
 ans  =
  F
```

2.2.1 Representação de números inteiros

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

Representação sem signal

Um registro com n bits da forma

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

representa o número $(d_{n-1}d_{n-2}...d_1d_0)_2$. Assim é possível representar números inteiros entre

$$\begin{aligned}(111...111)_2 &= 2^{n-1}+2^{n-2}+\cdots+2^1+2^0 = 2^n-1.\\ \vdots &= \\ (000...000)_2 &= 0\end{aligned}$$

Observação 2. *No Scilab, consulte sobre os comandos:* `uint8`, `uint16` *e* `uint32`.

Representação com bit de sinal

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

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

representa o número $(-1)^s(d_{n-2}...d_1d_0)_2$. Assim é possível representar números inteiros entre -2^{n-1} e 2^{n-1}, com duas representações para o zero: $(1000...000)_2$ e $(00000...000)_2$.

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

$$\begin{aligned}(11111111)_2 &= -(2^6+\cdots+2+1) = -127\\ &\vdots\\ (10000001)_2 &= -1\\ (10000000)_2 &= -0\\ (01111111)_2 &= 2^6+\cdots+2+1 = 127\\ &\vdots\\ (00000010)_2 &= 2\\ (00000001)_2 &= 1\\ (00000000)_2 &= 0\end{aligned}$$

Representação complemento de dois

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

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

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

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

Exemplo 9. *O registro com* 8 *bits* [01000011] *representa o número* $-0(2^7) + (1000011)_2 = 64 + 2 + 1 = 67$.

O registro com 8 *bits* [10111101] *representa o número* $-1(2^7)+(0111101)_2 = -128 + 32 + 16 + 8 + 4 + 1 = -67$.

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

Exemplo 10. *Em um registro com* 8 *bits, teremos os números*

$$
\begin{aligned}
(11111111)_2 &= -2^7 + 2^6 + \cdots + 2 + 1 = -1\\
&\vdots\\
(10000001)_2 &= -2^7 + 1 = -127\\
(10000000)_2 &= -2^7 = -128\\
(01111111)_2 &= 2^6 + \cdots + 2 + 1 = 127\\
&\vdots\\
(00000010)_2 &= 2\\
(00000001)_2 &= 1\\
(00000000)_2 &= 0
\end{aligned}
$$

Observação 3. *No Scilab, consulte sobre os comandos:* `int8`, `int16` *e* `int32`.

2.2.2 Sistema de ponto fixo

O sistema de ponto fixo representa as partes inteira e fracionária do número com uma quantidade fixas de dígitos. Por exemplo, em um computador de 32 bits que usa o sistema de ponto fixo, o registro

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

pode representar o número

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

  10000000000000000000000000000000

  e

  00000000000000000000000000000000

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

  11111111111111111111111111111111

  e

  00000000000000000000000000000000

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

  00000000000000000000000000000000

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

2.2.3 Sistema de ponto flutuante

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

Exemplo 11. *Um computador de 64 bits que usa o sistema de ponto flutuante com um dígito para o sinal, o registro:*

s	c_{10}	c_9	$\cdots$	c_0	m_{-1}	m_{-2}	$\cdots$	m_{-50}	m_{-51}

representa o número

$$(-1)^s 2^{c-1023}(1+m),$$

onde:

$$c = c_{10}2^{10} + c_9 2^9 + \cdots + c_1 2^1 + c_0 2^0$$

e

$$m = m_{-1}2^{-1} + m_{-2}2^{-2} + \cdots + m_{-50}2^{-50} + m_{-51}2^{-51}.$$

Os dígitos $c_{10}, c_9 \cdots, c_0$ *compõem a característica e* $m_{-1}, m_{-2} \cdots, m_{-51}$ *compõem a mantissa. Por exemplo, o número:*

0 10000000000 1000000000000000000000000000000000000000000000000000

representa o número:

$$(-1)^0 2^{1024-1023}(1 + 2^{-1}) = 3$$

O menor número positivo nesse sistema acontece quando $s = 0$, $c = 1$ *e* $m = 0$, *ou seja:*

$$2^{-1022}(1 + 0) \approx 0,2225 \times 10^{-307}.$$

O maior número acontece quando $s = 0$, $c = 2046$ *e* $m = 1 - 2^{-52}$, *ou seja:*

$$2^{1023}(2 - 2^{-52}) \approx 0,17977 \times 10^{309}.$$

$\Diamond$

Observação 4. *O menor número positivo e o maior número positivo representados no Scilab são, respectivamente:*

```
-->number_properties('tiny')
-->number_properties('huge')
```

Um número muito pequeno (**underflow**) geralmente é aproximado por zero e um número muito grande (**overflow**) geralmente faz o cálculo parar. Os demais números são aproximados pelo mais próximo, gerando os erros de arredondamento.

Observação 5. *O chamado modo de exceção de ponto flutuante é controlado pela função* `ieee`. *O padrão do Scilab é* `ieee(0)`. *Estude os seguintes resultados das seguintes operações usando os diferentes modos de exceção:*

```
-->2*number_properties('huge'), 1/2^999, 1/0, 1/-0
```

Definição 2. *(Forma normalizada de ponto flutuante) Seja* x *um número real não nulo.*[2] *A representação de* x *na forma normalizada em ponto flutuante na base* $b \in \mathbb{N}$, $b > 1$, *e com* n **dígitos significativos**[3] *é dada por:*

$$fl(x) = 0, d_1 d_2 d_3 \ldots d_n \times b^E$$

onde, $d_i \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$ *e* $E \in \mathbb{Z}$. *Os dígitos* d_i, $i = 1, 2, \ldots, n$, *formam o significando e* E *o expoente da representação de* x.

[2]Se $x = 0$, então $fl(x) = 0,000 \ldots 0 \times b^0$ será considerada sua representação em ponto flutuante.

[3]Na literatura, também encontra-se outros significados para o termo dígitos (algarismos) significativos. Veja, por exemplo, [3]

Exemplo 12. *O número* $0,05$ *é representado na forma normalizada de ponto flutuante na base* 2 *e com um dígito significativo por* $0,1 \times 2^{-1}$.

Observação 6. *Salvo especificado ao contrário, quando nos referirmos à representação em ponto flutuante de um número dado, estaremos nos referindo à representação deste número na forma normalizada de ponto flutuante na base dez.*

Exercício 4. *Represente os números* $0,00\overline{51}$ *e* $1205,41\overline{54}$ *em um sistema de ponto fixo de 4 dígitos para a parte inteira e 4 dígitos para a parte fracionária. Depois represente os mesmos números num sistema de ponto flutuante com 7 dígitos significativos.*

Solution. As representações dos números $0,00\overline{51}$ e $1205,41\overline{54}$ no sistema de ponto fixo são $0,0051$ e $1205,4154$, respectivamente. No sistema de ponto flutuante, as representações são $0,5151515 \cdot 10^{-2}$ e $0,1205415 \cdot 10^{4}$, respectivamente. ◊

Observação 7. *Consulte sobre o comando* `format` *no Scilab.*

2.3 Origem e Definição de Erros

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

1. Dados de entrada: equipamentos de medição possuem precisão finita, acarretando erros nas medidas físicas.

2. Erros de Truncamento: ocorrem quando aproximamos um procedimento formado por uma sequência infinita de passos através de um outro procedimento finito. Por exemplo, a definição de integral é dada por uma soma infinita e, como veremos na terceira área, aproximarmo-la por um soma finita. Esse é um assunto que discutiremos várias vezes no curso, pois o tratamento do erro de truncamento é feito para cada método numérico.

3. Erros de Arredondamento: são aqueles relacionados com as limitações que existem na forma representar números de máquina. Sobre esse tópico dedicamos a subseção (2.3.1).

Definição 3. *Seja* x *um número real e* $\overline{x}$ *sua aproximação. O* **erro absoluto** *da aproximação* $\overline{x}$ *é definido como sendo o número:*

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

O **erro relativo** *da aproximação $\overline{x}$ é definido como sendo o número:*

$$\frac{|x-\overline{x}|}{|x|}.$$

Observação 8. *Observe que o erro relativo é adimensional e, muitas vezes, é dado em porcentagem. Ou seja, o erro relativo, em porcentagem, da aproximação $\overline{x}$ é definido por:*

$$\frac{|x-\overline{x}|}{|x|}\times 100\%.$$

Exemplo 13. *Se $x=\frac{1}{3}$ e $\overline{x}=0,333$, então o erro absoluto é*

$$|x-\overline{x}|=|0,\overline{3}-0,333|=0,000\overline{3}=0,\overline{3}\cdot 10^{-3}$$

e o erro relativo é

$$\frac{|x-\overline{x}|}{|x|}=\frac{0,\overline{3}\cdot 10^{-3}}{0,\overline{3}}=10^{-3}=0,1\%$$

Exemplo 14. *Observe os erros absolutos e relativos em cada caso*

	erro absoluto	*erro relativo*
$x=0,\overline{3}\cdot 10^{-2}$ *e* $\overline{x}=0,3\cdot 10^{-2}$	$0,\overline{3}\cdot 10^{-3}$	$\frac{0,\overline{3}\cdot 10^{-3}}{0,\overline{3}\cdot 10^{-2}}=10^{-1}=10\%$
$x=0,\overline{3}$ *e* $\overline{x}=0,3$	$0,\overline{3}\cdot 10^{-1}$	$\frac{0,\overline{3}\cdot 10^{-1}}{0,\overline{3}}=10^{-1}=10\%$
$x=0,\overline{3}\cdot 10^{2}$ *e* $\overline{x}=0,3\cdot 10^{2}$	$0,\overline{3}\cdot 10^{1}$	$\frac{0,\overline{3}\cdot 10^{1}}{0,\overline{3}\cdot 10^{2}}=10^{-1}=10\%$

Exercício 5. *Calcule os erros absoluto e relativo das aproximações $\overline{x}$ para x*

a) $x=\pi=3,14159265358979\cdots$ *e* $\overline{x}=3,141$

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

c) $x=100001$ *e* $\overline{x}=100000$

Definição 4. *A aproximação $\overline{x}$ de um número $x=\pm 0,d_{-1}d_{-2}d_{-3}\cdots\times 10^{m}$ possui s* **dígitos significativos corretos** *se o erro absoluto $|x-\overline{x}|$ satisfazer*[4]

$$|x-\overline{x}|\leq 0,5\times 10^{m-s}$$

[4]Observação: Não existe uma definição única na literatura para o conceito de dígitos significativos corretos, embora não precisamente equivalentes, transmitem a mesmo conceito.

Exemplo 15. *Veja os seguintes casos:*

a) *Considere* $x = 0,\overline{3}$, $\overline{x} = 0,333$ *e o erro absoluto* $\delta = |x - \overline{x}| = 0,\overline{3} \times 10^{-3} = 0,\overline{3} \times 10^{0-3}$. *Essa aproximação tem 3 dígitos significativos corretos.*

b) *Agora, considere* $x = 10,00\overline{1} = 0,1000\overline{1} \times 10^2$, $\overline{x} = 9,99933 = 0,999933 \times 10^1$ *e o erro absoluto* $\delta = |x - \overline{x}| = 0,178\overline{1} \times 10^{-2} = 0,178\overline{1} \times 10^{2-4}$. *Essa aproximação possui todos os dígitos diferentes se comparamos um a um, mas tem 4 dígitos significativos corretos.*

Exercício 6. *Verifique quantos são os dígitos significativos corretos em cada aproximação* $\overline{x}$ *para* x.

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

b) $x = 100$ *e* $\overline{x} = 99$

2.3.1 Erros de Arredondamento

Os erros de arredondamento são aqueles gerados quando aproximamos um número real por um número com representação finita.

Exemplo 16. *O número* $\frac{1}{3} = 0,\overline{3}$ *possui um representação infinita tanto na base decimal quanto na base binária. Logo, quando representamos ele no computador geramos um erro de arredondamento que denotaremos por* ϵ. *Agora considere a seguinte sequência:*

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

Observe que $x_0 = \frac{1}{3}$, $x_1 = 4 \cdot \frac{1}{3} - 1 = \frac{1}{3}$, $x_2 = \frac{1}{3}$, *ou seja, temos uma sequência constante igual a* $\frac{1}{3}$. *Se calcularmos no computador essa sequência, temos que incluir os erros de arredondamento, ou seja,*

$$\begin{aligned} \tilde{x}_0 &= \frac{1}{3} + \epsilon \\ \tilde{x}_1 &= 4x_0 - 1 = 4\left(\frac{1}{3} + \epsilon\right) - 1 = \frac{1}{3} + 4\epsilon \\ \tilde{x}_2 &= 4x_1 - 1 = 4\left(\frac{1}{3} + 4\epsilon\right) - 1 = \frac{1}{3} + 4^2\epsilon \\ &\vdots \\ \tilde{x}_n &= \frac{1}{3} + 4^n\epsilon \end{aligned}$$

Portanto o limite da sequência diverge,

$$\lim_{x\to\infty} |\tilde{x}_n| = \infty$$

Faça o teste no scilab, colocando:

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

e itere algumas vezes a linha de comando:

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

Existem várias formas de aproximar um número em ponto flutuante $\pm 0, d_1 d_2 d_3 \ldots d_{k-1} d_k d_{k+1} \ldots d_n$ usando k dígitos significativos. As duas principais são as seguintes:

1. Por truncamento: aproximamos o número dado por:

$$\pm 0, d_1 d_2 d_3 \ldots d_k \times 10^e$$

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

2. Por arredondamento: aproximamos o número dado por:

$$\pm 0, \tilde{d}_1 \tilde{d}_2 \tilde{d}_3 \ldots \tilde{d}_k \times 10^{\tilde{e}}$$

que é a aproximação por truncamento do número:

$$\pm 0, d_1 d_2 d_3 \ldots d_k d_{k+1} \times 10^e \pm 0,5 \times 10^{e-k}$$

Exemplo 17. *Represente os números* $0,567$*;* $0,233$*;* $-0,6785$ *e* $\pi = 0,314159265... \times 10^1$ *com dois dígitos significativos por truncamento e arredondamento.*
Truncamento: $0,56$*;* $0,23$*;* $-0,67$ *e* $\pi = 0,31 \times 10^1 = 3,1$
Arredondamento: $0,57$*;* $0,23$*;* $-0,68$ *e* $\pi = 0,31 \times 10^1 = 3,1$

Observação 9. *Observe que o arredondamento pode mudar todos os dígitos e o expoente da representação em ponto flutuante de um número dado.*

Exemplo 18. *O arredondamento de* $0,9999 \times 10^{-1}$ *com* 3 *dígitos significatios é* $0,1 \times 10^0$.

Exercício 7. *Represente os números* 3276*;* $42,55$ *e* $0,00003331$ *com três dígitos significativos por truncamento e arredondamento.*

Exercício 8. *Resolva a equação* $0,1x - 0,01 = 12$ *usando arredondamento com três dígitos significativos em cada passo e compare com o resultado analítico*

2.4 Propagação de Erros

Dado uma função diferenciável f, considere $\overline{x}$ uma aproximação para x e $f(\overline{x})$ uma aproximação para $f(x)$. Sabendo o erro $\delta_x = |x - \overline{x}|$, queremos estimar o erro $\delta_f = |f(x) - f(\overline{x})|$. Pelo teorema do valor médio, existe ϵ contido no intervalo aberto formado por x e $\overline{x}$ tal que

$$f(x) - f(\overline{x}) = f'(\epsilon)(x - \overline{x}).$$

Como não conhecemos o valor de ϵ, supomos que a derivada $f'(\epsilon)$ é limitada por M ($|f'(\epsilon)| \leq M$) no intervalo fechado formado por x e $\overline{x}$ e obtemos

$$|f(x) - f(\overline{x})| \leq M|x - \overline{x}|.$$

Se $f'(x)$ não varia muito rápido nesse intervalo e supondo δ_x pequeno, aproximamos $M \approx |f'(x)|$ e temos:

$$|f(x) - f(\overline{x})| \approx |f'(x)||x - \overline{x}|,$$

ou

$$\delta_f \approx |f'(x)|\delta_x.$$

De modo geral, quando f depende de várias variáveis, a seguinte estimativa vale:

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

Exercício Resolvido 1. *Seja $f(x) = x\exp(x)$. Calcule o erro absoluto em se calcular $f(x)$ sabendo que $x = 2 \pm 0,05$.*

Solution. Temos que $x \approx 2$ com erro absoluto de $\delta_x = 0,05$. Neste caso, calculamos δ_f, i.e. o erro absoluto em se calcular $f(x)$, por:

$$\delta_f = |f'(x)|\delta_x.$$

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

$$\begin{aligned} \delta_f &= |(1 + x)e^x| \cdot \delta_x \\ &= |3e^2| \cdot 0,05 = 1,084. \end{aligned}$$

Portanto, o erro absoluto em se calcular $f(x)$ quando $x = 2 \pm 0,05$ é de $1,084$. ◊

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

Solution. Calculamos as derivadas parciais de f:

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

e

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

Calculamos o erro absoluto em termos do erro relativo:

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

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

Aplicando a expressão para estimar o erro em f temos

$$\begin{aligned}\delta_f &= \left|\frac{\partial f}{\partial x}\right|\delta_x + \left|\frac{\partial f}{\partial y}\right|\delta_y \\ &= \frac{2e^4}{27} \cdot 0,3 + 2\frac{9+1}{9}e^4 \cdot 0,06 = 8,493045557\end{aligned}$$

Portanto, o erro relativo ao calcular f é estimado por

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

$\Diamond$

Exercício Resolvido 3. *No exemplo anterior, reduza o erro relativo em* x *pela metade e calcule o erro relativo em* f*. Depois, repita o processo reduzindo o erro relativo em* y *pela metade.*

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

Exercício Resolvido 4. *Considere um triângulo retângulo onde a hipotenusa e um dos catetos são conhecidos a menos de um erro: hipotenusa* $a = 3 \pm 0,01$ *metros e cateto* $b = 2 \pm 0,01$ *metros. Calcule o erro absoluto ao calcular a área dessa triângulo.*

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

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

Agora calculamos as derivadas

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

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

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

$$\begin{aligned} \delta_A &\approx \left|\frac{\partial A}{\partial a}\right| \delta_a + \left|\frac{\partial A}{\partial b}\right| \delta_b \\ &\approx \frac{3\sqrt{5}}{5} \cdot 0,01 + \frac{\sqrt{5}}{10} \cdot 0,01 = 0,01565247584 \end{aligned}$$

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

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

$\Diamond$

Exercício 9. *A corrente* I *em ampères e a tensão* V *em volts em uma lâmpada se relacionam conforme a seguinte expressão:*

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

Onde α *é um número entre 0 e 1 e* V_0 *é a tensão nominal em volts. Sabendo que* $V_0 = 220 \pm 3\%$ *e* $\alpha = 0,8 \pm 4\%$ *Calcule a corrente e o erro relativo associado quando a tensão vale* $220 \pm 1\%$. **Dica:** *lembre que* $x^{\alpha} = e^{\alpha \ln(x)}$

2.5 Cancelamento Catastrófico

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

Exemplo 19. *Efetue a operação*

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

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

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

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

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

$$\frac{|0,987624687925 - 0,987625|}{|0,987624687925|} = 0,00003159\% \qquad e \qquad \frac{0,987624 - 0,987624}{0,987624} = 0\%,$$

enquanto no lado direito o erro relativo é enorme,

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

Exemplo 20. *Considere o problema de encontrar as raízes da equação de segundo grau:*

$$x^2 + 300x - 0,014 = 0,$$

usando seis dígitos significativos.

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

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

e as raízes:

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

Então, as duas raízes são:

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

e

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

Agora, os valores das raízes com seis dígitos significativos deveriam ser

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

Observe que um raiz saiu com seis dígitos significativos corretos, mas a outra não possui nenhum dígito significativo correto.

Observação 10. *No exemplo anterior b^2 é muito maior que $4ac$, ou seja, $b \approx \sqrt{b^2 - 4ac}$, logo a diferença*

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

estará próxima de zero. Uma maneira padrão de evitar o cancelamento catastrófico é usar procedimentos analíticos para eliminar essa diferença. Abaixo veremos alguns exemplos.

Exemplo 21. *Para eliminar o cancelamento catastrófico do exemplo anterior, usamos a seguinte expansão em série de Taylor em torno da origem*

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

Substituindo na fórmula de Bhaskara, temos:

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

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

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

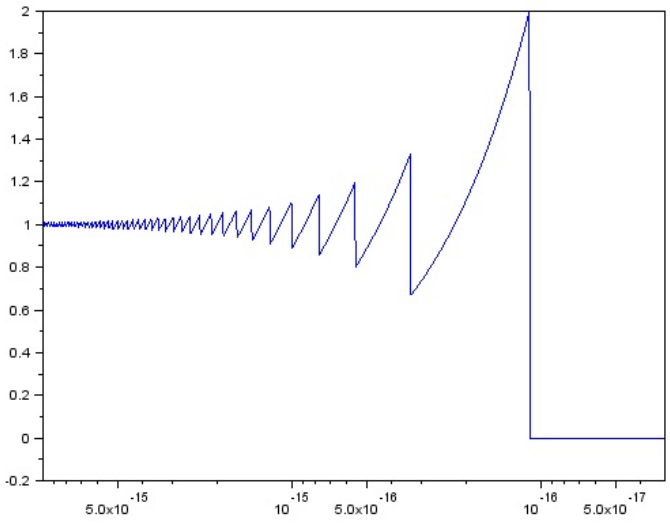
$$\begin{aligned} \tilde{x}_2 &= \frac{-b + b - \frac{4ac}{2b}}{2a} \\ &= -\frac{4ac}{4ab} \\ &= -\frac{c}{b} = -\frac{-0,140000 \times 10^{-1}}{0,300000 \times 10^3} = 0,466667 \times 10^{-4} \end{aligned}$$

Observe que o efeito catastrófico foi eliminado.

Exemplo 22. *Observe a seguinte identidade*

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

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



Exercício 10. *Considere a expressão*

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

para x pequeno. Verifique que

$$\lim_{x \to 0} f(x) = 0,5$$

Depois calcule no scilab $f(x)$ para $x = 10^{-5}$, $x = 10^{-6}$, $x = 10^{-7}$, $x = 10^{-8}$, $x = 10^{-9}$ e $x = 10^{-10}$. Finalmente, faça uma aproximação analítica que elimine o efeito catastrófico.

Exemplo 23. *Neste exemplo, estamos interessados em compreender mais detalhadamente o comportamento da expressão*

$$\left(1 + \frac{1}{n}\right)^n \tag{2.1}$$

quando n é um número grande ao computá-la em sistemas de numeral de ponto flutuante com acurácia finita. Um resultado bem conhecido do cálculo nos diz que o limite de (2.1) quando n tende a infinito é o número de Euler:

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = e = 2,718281828459... \tag{2.2}$$

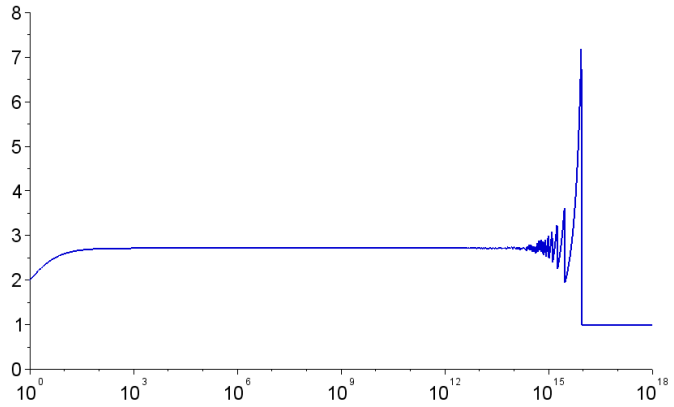
Sabemos também que a sequência produzida por (2.1) é crescente, isto é:

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

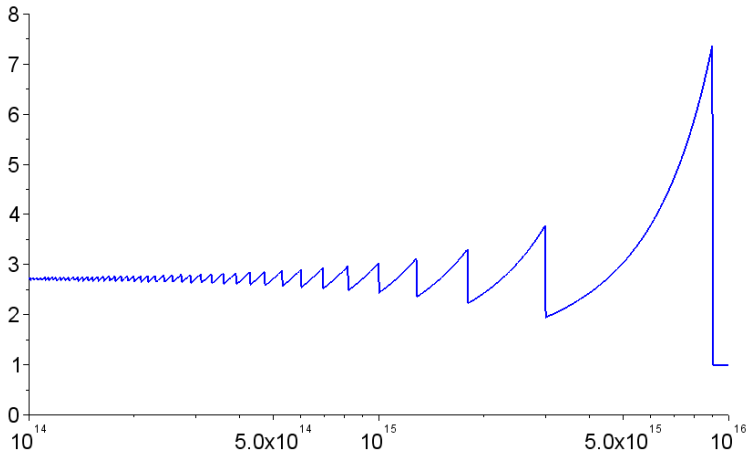
No entanto, quando calculamos essa expressão no Scilab, nos defrontamos com o seguinte resultado:

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

Podemos resumir esses dados no seguinte gráfico de $\left(1+\frac{1}{n}\right)^n$ *em função de* n*:*



Observe que quando x *se torna grande, da ordem de* 10^{15}*, o gráfico da função deixa de se crescente e apresenta oscilações. Observe também que a expressão se torna identicamente igual a* 1 *depois de um certo limiar. Tais fenômenos não são intrínsecos da função* $f(x) = (1+1/x)^x$*, mas* oriundas de erros de arredondamento*, isto é, são resultados numéricos espúrios. A fim de pôr o comportamento numérico de tal expressão, apresentamos abaixo o gráfico da mesma função, porém restrito à região entre* 10^{14} *e* 10^{16}*.*



Para compreender por que existe um limiar N que, quando atingido torna a expressão identicamente igual a 1, observe a sequência de operações realizadas pelo computador:

$$x \;\rightarrow\; 1/x \;\rightarrow\; 1+1/x \;\rightarrow\; (1+1/x)^x \tag{2.3}$$

Devido ao limite de precisão da representação de números em ponto flutuante, existe um menor número representável que é maior do que 1. Este número pode ser obtido pelo comando:

```
-->1+%eps
 ans  =
    1.0000000000000002220446
```

A quantidade dada por `%eps` *é chamada de* **épsilon de máquina** *e é o menor número que somado a 1 produz um resultado superior a 1 no sistema de numeração usado. O épsilon de máquina no sistema de numeração "double" vale aproximadamente $2,22 \times 10^{-16}$. Quando somamos a 1 um número positivo inferior ao épsilon de máquina, obtemos o número 1. Dessa forma, o resultado obtido pela operação de ponto flutuante $1+x$ para $0 < x < 2,22 \times 10^{-16}$ é 1.*

Portanto, quando realizamos a sequência de operações dada em (2.3), toda informação contida no número x é perdida na soma com 1 quando $1/x$ é menor que o épsilon de máquina, o que ocorre quando $x > 5 \times 10^{15}$. Assim $(1+1/x)$ é aproximado para 1 e a última operação se resume a 1^x, o que é igual a 1 mesmo quando x é grande.

Um erro comum é acreditar que o perda de significância se deve ao fato de $1/x$ ser muito pequeno para ser representado e é aproximando para 0.

Isto é falso, o sistema de ponto de flutuante permite representar números de magnitude muito inferior ao épsilon de máquina. O problema surge da limitação no tamanho da mantissa. Observe como a seguinte sequência de operações não perde significância para números positivos x muito menores que o épsilon de máquina:

$$x \ \rightarrow \ 1/x \ \rightarrow \ 1/(1/x) \tag{2.4}$$

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

$$x \ \rightarrow \ 1+x \ \rightarrow \ (1+x)-1. \tag{2.5}$$

Finalmente, notamos que quando tentamos calcular $\left(1+\frac{1}{n}\right)^n$ para n grande, existe perda de significância no cálculo de $1+1/n$. Para entender isso, observe o que acontece quando $n = 7 \times 10^{13}$:

```
-->n=7e13
 n  =
     7.000000000000000000D+13

-->1/n
 ans  =
     1.428571428571428435D-14

-->y=1+1/n
 y  =

    1.000000000000014211D+00
```

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

```
-->y-1
 ans  =
     1.421085471520200372D-14

-->1/(y-1)
 ans  =
     7.036874417766400000D+13
```

Exemplo 24 (Analogia da balança)**.** *Observe a seguinte comparação interessante que pode ser feita para ilustrar os sistemas de numeração com ponto fixo e flutuante: o sistema de ponto fixo é como uma balança cujas marcas estão igualmente espaçadas; o sistema de ponto flutuante é como uma balança cuja distância entre as marcas é proporcional à massa medida. Assim, podemos ter uma balança de ponto fixo cujas marcas estão sempre distanciadas de* $100g$ *(*$100g$*,* $200g$*,* $300g$*, ...,* $1Kg$*,* $1,1Kg$*,...) e outra balança de ponto flutuante cujas marcas estão distanciadas sempre de aproximadamente um décimo do valor lido (*$100g$*,* $110g$*,* $121g$*,* $133g$*, ...,* $1Kg$*,* $1,1Kg$*,* $1,21Kg$*, ...) A balança de ponto fixo apresenta uma resolução baixa para pequenas medidas, porém uma resolução alta para grandes medidas. A balança de ponto flutuante distribui a resolução de forma proporcional ao longo da escala.*

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

Capítulo 3

Solução de equações de uma variável

Neste capítulo buscaremos aproximações numéricas para raízes de funções de uma variável que são continuamente diferenciáveis.

3.1 Condição de Existência de raízes reais

Um teorema que garante a existência de raiz real em um intervalo é o teorema do valor intermediário:

Teorema 1 (Teorema do Valor Intermediário)**.** *Se $f : [a, b] \to \mathbb{R}$ é um função contínua e K for um número entre $f(a)$ e $f(b)$, então existe $c \in (a, b)$ para o qual $f(c) = K$.*

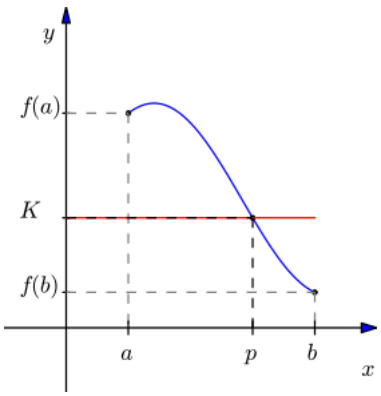


Figura 3.1: Teorema do valor intermediário

Em particular, se $f(a) > 0$ e $f(b) < 0$, então $0 \in [f(b), f(a)]$ e podemos garantir a existência de $c \in (a, b)$ tal que $f(c) = 0$, i.e. existe uma raiz no intervalo (a, b). A mesma afirmação é válida se $f(a) < 0$ e $f(b) > 0$.

Exemplo 25. *Mostre que existe pelo menos uma solução da equação* $e^x = x + 2$ *no intervalo* $(-2, 0)$.

De fato, se tomarmos $f(x) = e^x - x - 2$, *então* $f(0) = 1 - 2 < 0$ *e* $f(-2) = e^{-2} + 2 - 2 > 0$. *Pelo teorema do valor intermediário, existe* $c \in (-2, 0)$ *tal que* $f(c) = 0$, *ou seja, existe pelo menos uma solução nesse intervalo.*

Quando procuramos aproximações para raízes de funções, é importante que cada um delas fique isolada em um intervalo. Ou seja, precisamos garantir a existência e a unicidade da raiz. A existência vem do teorema do valor intermediário e a unicidade vem da monotonicidade da função.

Teorema 2. *Se* $f : [a, b] \to \mathbb{R}$ *é um função diferenciável,* $f(a) \cdot f(b) < 0$ *e* $f'(x) > 0$ *(ou* $f'(x) < 0$*) para* $x \in (a, b)$, *então existe uma única raiz* c *em* (a, b).

Em outras palavras, se a função corta o eixo x e é sempre crescente (ou sempre decrescente), então a raiz é única.

Exemplo 26. *Observamos que existe uma única solução da equação* $e^x = x + 2$ *no intervalo* $(-2, 0)$. *A existência foi estabelecida no exemplo anterior. Para garantir a unicidade, observe que* $f'(x) = e^x - 1$ *e, portanto,* $f'(x) < 0$ *para* $x \in (-2, 0)$. *Logo a raiz é única.*

Podemos inspecionar o comportamento da função $f(x)e^x - x - 2$ *e de sua derivada fazendo seus gráficos no Scilab. Para tanto, podemos implementar o seguinte código:*

```
-->x = linspace(-2,0,50);
-->//grafico de f(x)
-->deff('y = f(x)','y=exp(x)-x-2')
-->plot(x,f(x))
-->//graficando a f'(x)
-->deff('y = fl(x)','y=exp(x)-1')
-->plot(x,fl(x))
```

Exercício 11. *Mostre que a equação*

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

possui uma única solução positiva. Faça o gráfico e observe.

Exercício 12. *Use o teorema do valor intermediário para mostrar que o erro absoluto ao aproximar a raiz da função* $f(x) = e^x - x - 2$ *por* $\overline{x} = -1,841$ *é menor que* 10^{-3}.

3.2 Método da bisseção

Suponha que a função contínua $f : [a, b] \to \mathbb{R}$ tal que $f(a) \cdot f(b) < 0$, ou seja, f possui uma raiz no intervalo. Suponha também que a raiz é única. Uma primeira aproximação para a raiz pode ser o ponto médio $p = \frac{a+b}{2}$. Se $f(p) \cdot f(a) < 0$, então a raiz está a esquerda de p, se não, a raiz está a direita de p (veja Fig. 3.2). Depois de escolher o intervalo correto, fazemos uma nova aproximação para a raiz tomando o ponto médio do novo intervalo.

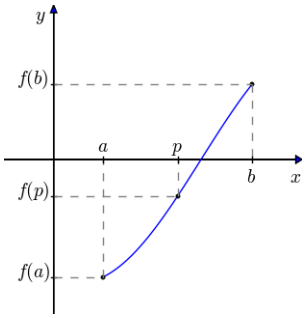


Figura 3.2: Método da bisseção.

Em outras palavras, seja $(a^{(0)}, b^{(0)}) = (a, b)$ o intervalo inicial e $p^{(0)} = \frac{a^{(0)}+b^{(0)}}{2}$ a aproximação inicial. Se $f(p^{(0)}) \cdot f(a^{(0)}) < 0$, então $(a^{(1)}, b^{(1)}) = (a^{(0)}, p^{(0)})$, caso contrário, $(a^{(1)}, b^{(1)}) = (p^{(0)}, b^{(0)})$. A nova aproximação para a raiz é $p^{(1)} = \frac{a^{(1)}+b^{(1)}}{2}$. Esse procedimento produz uma sequência $p^{(n)}$ que converge para a raiz.

Exemplo 27. *Faça 5 iterações do método da bisseção para encontrar a raiz de* $f(x) = x^3 + 5x^2 - 12$ *utilizando* $a^{(0)} = 1$ *e* $b^{(0)} = 2$.

n	$a^{(n)}$	$b^{(n)}$	$p^{(n)}$	$f(a^{(n)})$	$f(b^{(n)})$	$f(p^{(n)})$
0	$a^{(0)} = 1$	$b^{(0)} = 2$	$p^{(0)} = 1,5$	-6	16	$2,625$
1	$a^{(1)} = 1$	$b^{(1)} = p^1 = 1,5$	$p^{(1)} = 1,25$	-6	$2,625$	$-2,234375$
2	$a^{(2)} = 1,25$	$b^{(2)} = 1,5$	$p^{(2)} = 1,375$			
3						
4						
5						

No console do Scilab, temos:

```
-->deff('y=f(x)','y = x^3 + 5*x^2 - 12')
-->//iteracao 0
-->a=1; b=2; p=(a+b)/2;
-->[a,b,p,f(a),f(b),f(p)]
 ans  =
    1.    2.    1.5  - 6.    16.    2.625
-->//iteracao 1
-->b = p; p = (a+b)/2;
-->[a,b,p,f(a),f(b),f(p)]
 ans  =
    1.    1.5    1.25  - 6.    2.625  - 2.234375
```

Observe que a distância entre $p^{(0)}$ e a raiz p^* não pode exceder metade do intervalo, ou seja $|p^{(0)} - p^*| \leq \frac{b-a}{2}$. Da mesma forma, o erro absoluto entre $p^{(1)}$ e p^* é menor que $\frac{1}{4}$ do intervalo, isto é, $|p^{(1)} - p^*| \leq \frac{b-a}{2^2}$. De modo geral, o erro absoluto na iteração n é estimado por

$$|p^{(n)} - p^*| \leq \frac{b-a}{2^{n+1}}, \qquad n \geq 1.$$

Também, se $\epsilon_n := |p^{(n)} - p^*|$, então vale:

$$\epsilon_{n+1} \leq \frac{1}{2}\left(\epsilon_n\right)^1$$

e, por isso, dizemos que o método da bisseção possui taxa de convergência linear. Um método com taxa de convergência super-linear satisfaz

$$\epsilon_{n+1} \leq C\left(\epsilon_n\right)^m,$$

onde $m > 1$ e C é uma constante.

Exemplo 28. *Determine quantas iterações são necessárias para encontrar a raiz de* $f(x) = x^3 + 5x^2 - 12$ *com uma precisão de* 10^{-3}*, utilizando* $a^{(0)} = 1$ *e* $b^{(0)} = 2$*.*

Observe que precisamos da seguinte desigualdade

$$|p^{(n)} - p^*| \leq \frac{b-a}{2^{n+1}} = \frac{1}{2^{n+1}} \leq 10^{-3}.$$

Assim,

$$\log_2 2^{-(n+1)} \leq \log_2 10^{-3}$$

ou seja,

$$-(n+1)\log_2 2 \leq -3\log_2(10) \Rightarrow n+1 \geq 3\log_2(10) \approx 9,97 \Rightarrow n \approx 8,97$$

Portanto, $n \geq 9$*.*

Exercício 13. *Utilize o método da bisseção na equação* $\sqrt{x} = \cos(x)$ *para encontrar* $p^{(4)}$ *em* $[a, b] = [0, 1]$*.*

3.2.1 Código Scilab

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

- `f` - função objetivo
- `a` - extremo esquerdo do intervalo de inspeção $[a, b]$
- `b` - extremo direito do intervalo de inspeção $[a, b]$
- `TOL` - tolerância (critério de parada)
- `N` - número máximo de iterações

A variável de saída é:

- `p` - aproximação da raiz de `f`, i.e. $f(p) \approx 0$.

```
function [p] = bissecao(f, a, b, TOL, N)
  i = 1
  fa = f(a)
  while (i <= N)
    //iteracao da bissecao
    p = a + (b-a)/2
    fp = f(p)
    //condicao de parada
    if ((fp == 0) | ((b-a)/2 < TOL)) then
      return p
    end
    //bissecta o intervalo
    i = i+1
    if (fa * fp > 0) then
      a = p
      fa = fp
    else
      b = p
    end
  end
  error ('Num. max. de iter. excedido!')
endfunction
```

Exercício 14. *Encontre a solução de cada equação com erro absoluto inferior a* 10^{-6}.

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

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

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

Exercício 15. *Encontre numericamente as três primeiras raízes positivas da equação dada por*

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

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

Exercício 16. *Calcule uma equação da reta tangente a curva* $y = e^{-(x-1)^2}$ *que passa pelo ponto* $(3, 1/2)$.

3.3 Iteração de Ponto Fixo

3.3.1 Exemplo Histórico

Vamos analisar o método babilônico para extração da raiz quadrada de um número positivo A usando operações de soma, subtração, divisão e multiplicação.

Seja $x > 0$ uma aproximação para $\sqrt{A}$, temos três caso:

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

É natural imaginar que uma melhor aproximação para $\sqrt{A}$ é dada por

$$y = \frac{x + \frac{A}{x}}{2}$$

Aplicando esse método repetidas vezes, construímos a seguinte iteração:

$$\begin{aligned} x^{(n+1)} &= \frac{x^{(n)}}{2} + \frac{A}{2x^{(n)}} \\ x^{(0)} &= x \end{aligned}$$

Exemplo 29. *A=5, x=2*

$$\begin{aligned} x^{(n+1)} &= \frac{x^{(n)}}{2} + \frac{2,5}{x^{(n)}} \\ x^{(0)} &= 2 \end{aligned}$$

$$\begin{aligned} x^{(0)} &= 2 \\ x^{(1)} &= \frac{2}{2} + \frac{2,5}{2} = 1 + 1,25 = 2,25 \\ x^{(2)} &= \frac{2,25}{2} + \frac{2,5}{2,25} = 2,2361111 \\ x^{(3)} &= \frac{2,2361111}{2} + \frac{2,5}{2,2361111} = 2,236068 \\ x^{(4)} &= \frac{2,236068}{2} + \frac{2,5}{2,236068} = 2,236068 \end{aligned}$$

Exemplo 30. *A=10, x=1*

$$\begin{aligned} x^{(n+1)} &= \frac{x^{(n)}}{2} + \frac{5}{x^{(n)}} \\ x^{(0)} &= 1 \end{aligned}$$

$$\begin{aligned} x^{(0)} &= 1 \\ x^{(1)} &= \frac{1}{2} + \frac{5}{1} = 0,5 + 5 = 5,5 \\ x^{(2)} &= \frac{5,5}{2} + \frac{5}{5,5} = 3,6590909 \\ x^{(3)} &= \frac{3,6590909}{2} + \frac{5}{3,6590909} = 3,1960051 \\ x^{(4)} &= \frac{3,1960051}{2} + \frac{5}{3,1960051} = 3,1624556 \\ x^{(5)} &= \frac{3,1624556}{2} + \frac{5}{3,1624556} = 3,1622777 \\ x^{(6)} &= \frac{3,1622777}{2} + \frac{5}{3,1622777} = 3,1622777 \end{aligned}$$

A experimentação numérica sugere que o método funciona, mas três perguntas devem ser respondidas:

1. Será que a sequência é convergente?

2. Caso seja convergente, será que o limite $x^* = \lim_{n\to\infty} x_n$ é igual a $\sqrt{A}$?

3. Caso seja convergente, quão rápida é a convergência?

A segunda pergunta é a mais fácil de ser respondida:
Supondo que o limite de x_n exista, basta substituir na iteração:

$$\begin{aligned}
\lim_{n\to\infty} x^{(n+1)} &= \lim_{n\to\infty} \frac{x^{(n)}}{2} + \lim_{n\to\infty} \frac{A}{2x^{(n)}} \\
x^* &= \frac{x^*}{2} + \frac{A}{2x^*} \\
\frac{x^*}{2} &= \frac{A}{2x^*} \\
x^* &= \frac{A}{x^*} \\
(x^*)^2 &= A \\
x^* &= \sqrt{A}
\end{aligned}$$

Portanto, sempre que esse método converge, temos a garantia de que o limite é $\sqrt{A}$. (Independente do valor inicial!)

De fato, podemos provar que o método é convergente para qualquer valor inicial positivo x. E, ainda, que a convergência é rápida (ainda precisamos definir isso).

Para responder essas perguntas, devemos formalizar o conceito de ponto fixo. Antes disso, analisemos mais um exemplo:

3.3.2 Outro Exemplo

Suponha que queiramos resolver a equação:

$$xe^x = 10.$$

Observamos que o este problema é equivalente a resolver:

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

ou:

$$x = 10e^{-x}$$

Para tanto, vamos propor os seguintes processos iterativos:

$$a) \begin{cases} x^{(n+1)} = \ln\left(\frac{10}{x^{(n)}}\right), & n \geq 0 \\ x^{(0)} = 1 \end{cases}$$

e

$$b) \begin{cases} x^{(n+1)} = 10e^{-x^{(n)}}, & n \geq 0 \\ x^{(0)} = 1 \end{cases}$$

O processo $a)$ produz a seguinte sequência:

$$\begin{aligned} x^{(0)} &= 1 \\ x^{(1)} &= \ln(10) = 2,3025851 \\ x^{(2)} &= \ln\left(\frac{10}{2,3025851}\right) = 1,4685526 \\ x^{(3)} &= \ln\left(\frac{10}{1,4685526}\right) = 1,9183078 \\ x^{(4)} &= \ln\left(\frac{10}{1,9183078}\right) = 1,6511417 \\ &\vdots \\ x^{(10)} &= 1,7421335 \\ x^{(20)} &= 1,7455151 \\ x^{(30)} &= 1,745528 \\ x^{(31)} &= 1,745528 \end{aligned}$$

O processo $b)$ produz a seguinte sequência:

$$\begin{aligned} x^{(0)} &= 1 \\ x^{(1)} &= 10e^{-1} = 3,6787944 \\ x^{(2)} &= 10e^{-3,6787944} = 0,2525340 \\ x^{(3)} &= 10e^{-0,2525340} = 7,7682979 \\ x^{(4)} &= 10e^{-7,7682979} = 0,0042293 \\ x^{(5)} &= 10e^{-0,0042293} = 9,9577961 \end{aligned}$$

O experimento numérico sugere que o processo a não é convergente e que o processo b converge para $1,745528$.

3.3.3 Ponto fixo

Seja $\phi(x)$ uma função, dizemos que $x^* \in D(f)$ é um ponto fixo de ϕ se

$$\phi(x^*) = x^*$$

Seja $\phi : [a, b] \to [a, b]$ um função real tal que

$$|\phi(x) - \phi(y)| \leq \beta|x - y|, \quad \beta < 1.$$

Então ϕ é dita uma contração e existe um único ponto $x^* \in [a, b]$ tal que $\phi(x^*) = x^*$. Além disso, a sequência

$$x^{(n+1)} = \phi(x^{(n)})$$

é convergente sempre que $x_0 \in [a, b]$ e vale o limite

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

Observação 11. *A desigualdade $|\phi(x) - \phi(y)| \leq \beta|x - y|$ implica que $\phi(x)$ é contínua.*

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

$$f(a) = a - \phi(a) \leq a - a = 0$$

e

$$f(b) = b - \phi(b) \geq b - b = 0$$

Se $f(a) = a$ ou $f(b) = b$, então o ponto fixo existe. Caso contrário, as desigualdade são estritas e a função muda de sinal no intervalo. Como a função é contínua, pelo teorema do valor intermediário, existe um ponto x^* no intervalo (a, b) tal que $f(x^*) = 0$, ou seja, $x^* - \phi(x^*) = 0$. Observe que x^* é um ponto fixo de ϕ, pois $\phi(x^*) = x^*$.

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

$$|x^* - x^{**}| = |\phi(x^*) - \phi(x^{**})| \leq \beta|x^* - x^{**}|$$

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

Para demonstrar a convergência da sequência, observamos a seguinte relação

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

Agora observamos que

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

Portanto

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

e

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

Observações:

- A condição $|\phi(x) - \phi(y)| \leq \beta|x - y|$ é satisfeita sempre que $|\phi'(x)| \leq \beta < 1$ em todo o intervalo pois

$$|\phi(x)-\phi(y)| = \left|\int_x^y \phi'(s)ds\right| \leq \int_x^y |\phi'(s)|ds \leq \int_x^y \beta ds = \beta|x-y|, \;\; x < y.$$

- A desigualdade estrita $\beta < 1$ é necessária.
- A condição $f([a, b]) \subseteq [a, b]$ é necessária.

3.3.4 Teste de convergência

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

Teorema: Se $\phi \in C^1[a, b]$ e $|\phi'(x^*)| < 1$, então x^* é estável. Se $|\phi'(x^*)| > 1$ é instável e o teste é inconclusivo se $|\phi'(x^*)| = 1$.

Exemplo 31. *Considere o problema de encontrar a solução da equação algébrica*

$$\cos(x) = x$$

vendo-a como o ponto fixo da função

$$f(x) = \cos(x).$$

Mostraremos que o teorema do ponto fixo se aplica a esta função com $[a, b] = [1/2, 1]$.

Precisamos provar:

1. $f([1/2, 1]) \subseteq [1/2, 1]$;
2. $|f'(x)| < \beta, \quad \beta < 1, \quad \forall x \in [1/2, 1]$.

Para provar o item 1, observamos que $f(x)$ *é decrescente no intervalo, pelo que temos:*

$$0,54 < \cos(1) \leq \cos(x) \leq \cos(1/2) < 0,88$$

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

Para provar o item 2, observamos que

$$f'(x) = -\sin(x)$$

Da mesma forma, temos a estimativa:

$$-0,85 < -\sin(1) \leq -\sin(x) \leq -\sin(1/2) < -0,47$$

Assim, $|f'(x)| < 0,85$ *temos a desigualdade com* $\beta = 0,85 < 1$*.*

Agora, observamos o comportamento numérico da sequência:

$$\begin{cases} x^{(n+1)} = \cos(x^{(n)}), & n \geq 0 \\ x^{(0)} = 1 \end{cases}$$

Os primeiros termos podem ser calculados numericamente e são dados por:

$$\begin{aligned}
x^{(1)} &= \cos(x_0) = \cos(1) = 0,5403023 \\
x^{(2)} &= \cos(x_1) = \cos(0,5403023) = 0,8575532 \\
x^{(3)} &= \cos(x_2) = \cos(0,8575532) = 0,6542898 \\
x^{(4)} &= \cos(x_3) = \cos(0,6542898) = 0,7934804 \\
x^{(5)} &= \cos(x_4) = \cos(0,7934804) = 0,7013688 \\
x^{(6)} &= \cos(x_5) = \cos(0,7013688) = 0,7639597 \\
x^{(7)} &= \cos(x_6) = \cos(0,7639597) = 0,7221024 \\
x^{(8)} &= \cos(x_7) = \cos(0,7221024) = 0,7504178 \\
x^{(9)} &= \cos(x_8) = \cos(0,7504178) = 0,7314040 \\
x^{(10)} &= \cos(x_9) = \cos(0,7314040) = 0,7442374 \\
x^{(11)} &= \cos(x_{10}) = \cos(0,7442374) = 0,7356047 \\
x^{(12)} &= \cos(x_{11}) = \cos(0,7356047) = 0,7414251 \\
x^{(13)} &= \cos(x_{12}) = \cos(0,7414251) = 0,7375069 \\
&\vdots \\
x^{(41)} &= \cos(x_{40}) = \cos(0,7390852) = 0,7390851 \\
x^{(42)} &= \cos(x_{41}) = \cos(0,7390851) = 0,7390851 \\
x^{(43)} &= \cos(x_{42}) = \cos(0,7390851) = 0,7390851
\end{aligned}$$

Problema 1. *Resolver os problemas 33 e 34 da lista.*

3.3.5 Estabilidade e convergência

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

$$\begin{aligned} x^{(n+1)} &= \phi\left(x^{(n)}\right) \\ x^{(0)} &= x \end{aligned}$$

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

$$\begin{aligned} \phi(x) &= \phi(x^*) + (x - x^*)\phi'(x^*) + O\left((x - x^*)^2\right), n \geq 0 \\ &= x^* + (x - x^*)\phi'(x^*) + O\left((x - x^*)^2\right) \\ &\approx x^* + (x - x^*)\phi'(x^*) \end{aligned}$$

Substituindo na relação de recorrência, temos

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

Ou seja:

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

Tomando módulos, temos:

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

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

Conclusões:

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

Fixaremos, portanto, nos casos quando $|\phi'(x^*)| < 1$.

3.3.6 Erro absoluto e tolerância

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

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

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

Para responder a esta pergunta, observamos que

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

portanto:

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

Usamos também as expressões:

$$\begin{aligned} x^{(n+1)} &\approx x^* + (x^{(n)} - x^*)\phi'(x^*) \\ x^{(n)} &\approx x^* + (x^{(n-1)} - x^*)\phi'(x^*) \end{aligned}$$

Subtraindo uma da outra, temos:

$$x^{(n+1)} - x^{(n)} \approx (x^{(n)} - x^{(n-1)})\phi'(x^*)$$

Portanto:

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

E temos:

$$\begin{aligned} x^* - x^{(N)} &= \sum_{k=0}^{\infty} \left(x^{(N+k+1)} - x^{(N+k)}\right) \\ &\approx \sum_{k=0}^{\infty} \left(x^{(N+1)} - x^{(N)}\right)\left(\phi'(x^*)\right)^k \\ &= (x^{(N+1)} - x^{(N)})\frac{1}{1-\phi'(x^*)}, |\phi'(x^*)| < 1 \end{aligned}$$

Tomando módulo, temos:

$$\begin{aligned} \left|x^* - x^{(N)}\right| &\approx \left|x^{(N+1)} - x^{(N)}\right| \frac{1}{1-\phi'(x^*)} \\ \epsilon_N &\approx \frac{\Delta_N}{1-\phi'(x^*)} \end{aligned}$$

Conclusões: Tendo em mente a relação $x^{(n+1)} - x^{(n)} \approx (x^{(n)} - x^{(n-1)})\phi'(x^*)$, concluímos:

- Quando $\phi'(x^*) < 0$, o esquema é alternante e o erro ϵ_N pode ser estimado diretamente da diferença Δ_N.
- Quando $\phi'(x^*) > 0$, o esquema é monótono e $\frac{1}{1-\phi'(x^*)} > 1$, pelo que o erro ϵ_N é maior que a diferença Δ_N. A relação será tão mais importante quando mais próximo da unidade for $\phi'(x^*)$, ou seja, quando mais lenta for a convergência.
- Como $\phi'(x^*) \approx \frac{x^{(n+1)} - x^{(n)}}{x^{(n)} - x^{(n-1)}}$, temos

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

  e portanto

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

Observação 12. *Deve-se exigir que* $\Delta_n < \Delta_{n-1}$

Problema 2. *Resolver problemas 30 a 34 da lista.*

3.3.7 Problemas para análise

Problema 3. *Verifique (analiticamente) que a única solução real da equação*

$$xe^x = 10$$

é ponto fixo das seguintes funções:

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

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

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

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

Problema 4. *Verifique (analiticamente) que a única solução real da equação*

$$\cos(x) = x$$

é ponto fixo das seguintes funções:

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

b) $\phi(x) = 0,4x + 0,6\cos(x)$

c) $\phi(x) = x + \frac{\cos(x)-x}{1+\sin(x)}$

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

3.4 Método de Newton-Raphson

Consideramos o problema de encontrar as raízes da equação

$$f(x) = 0$$

onde $f(x) \in C^1$ através do método do ponto fixo. Para tal, observamos que um número real x^* é raiz de $f(x)$ se e somente se x^* é um ponto fixo da função

$$\phi(x) = x + \gamma(x)f(x), \quad \gamma(x) \neq 0$$

Aqui $\gamma(x)$ é uma função que será escolhida com base nos critérios de convergência do processo iterativo.

A derivada de $\phi(x)$ vale

$$\phi'(x) = 1 + \gamma(x)f'(x) + \gamma'(x)f(x)$$

no ponto x^*, temos

$$\phi'(x^*) = 1 + \gamma(x^*)f'(x^*) + \gamma'(x^*)f(x^*)$$

como $f(x^*) = 0$, temos

$$\phi'(x^*) = 1 + \gamma(x^*)f'(x^*)$$

Sabemos que o processo iterativo converge tão mais rápido quanto menor for $\phi'(x)$ nas vizinhanças de x^*, portanto, supomos que $f'(x^*) \neq 0$ e escolhemos $\gamma(x^*)$ de forma que

$$\phi'(x^*) = 0,$$

ou seja

$$\gamma(x^*) = -\frac{1}{f'(x^*)}.$$

Observe que x^* é raiz de $f(x)$ se, e somente se x^* é ponto fixo de

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

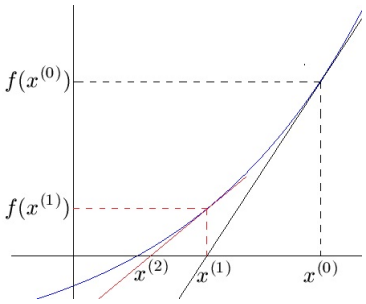
e $\phi'(x^*) = 0 < 1$. Portanto, o teorema do ponto fixo garante que se $x^{(0)}$ for suficientemente próximo a x^*, então o processo iterativo dado por

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}$$

converge para x^*, desde que $f'(x^{(n)}) \neq 0$ para todo $n \in \mathbb{N}$.

3.4.1 Interpretação Geométrica

Considere o problema de calcular a raiz uma função f, conforme esboço na figura abaixo



Queremos calcular $x^{(1)}$ em função de $x^{(0)}$ sabendo que é o corte da reta tangente em $x^{(0)}$ com o eixo x. A equação da reta que passa por $(x^{(0)}, f(x^{(0)})$ e é tangente a curva em $x^{(0)}$ tem inclinação $m = f'(x^{(0)})$ e sua equação é

$$y - f(x^{(0)}) = f'(x^{(0)})(x - x^{(0)}).$$

Sabendo que essa reta passa por $(x^{(1)}, 0)$, temos:

$$0 - f(x^{(0)}) = f'(x^{(0)})(x^{(1)} - x^{(0)}).$$

Portanto,

$$x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}$$

que é uma iteração do método de Newton. Repetimos o processo para calcular $x^{(2)},\ x^{(3)}, \ldots$. De modo geral, temos:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}.$$

Problema 5. *A partir do problema 23 da lista, encontre com pelo menos cinco dígitos significativos as três primeiras raízes da função $f(x)$.*

3.4.2 Análise de convergência

Seja $f(x)$ um função com derivada e derivada segunda contínuas tal que $f(x^*) = 0$ e $f'(x^*) \neq 0$. Seja também a função $\phi(x)$ definida como

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

Expandimos em série de Taylor em torno de x^* e obtermos:

$$\phi(x) = \phi(x^*) + (x - x^*)\phi'(x^*) + (x - x^*)^2 \frac{\phi''(x^*)}{2} + O\left((x - x^*)^3\right)$$

Sabemos que

$$\begin{aligned} \phi(x^*) &= x^* - \frac{f(x^*)}{f'(x^*)} = x^* \\ \phi'(x^*) &= 1 - \frac{f'(x^*)f'(x^*) - f(x^*)f''(x^*)}{\left(f'(x^*)\right)^2} = 1 - 1 = 0 \end{aligned}$$

Portanto:

$$\begin{aligned} \phi(x) &= x^* + (x - x^*)^2 \frac{\phi''(x^*)}{2} + O\left((x - x^*)^3\right) \\ &\approx x^* + (x - x^*)^2 \frac{\phi''(x^*)}{2}. \end{aligned}$$

Logo,

$$\begin{aligned} x^{(n+1)} &= \phi(x^{(n)}) \\ &\approx x^* + (x^{(n)} - x^*)^2 \frac{\phi''(x^*)}{2} \end{aligned}$$

$$\left(x^{(n+1)} - x^*\right) \approx (x^{(n)} - x^*)^2 \frac{\phi''(x^*)}{2}$$

Observação 13. *Pode-se mostrar facilmente que*

$$\phi''(x^*) = \frac{f''(x^*)}{f'(x^*)}$$

3.5 Método das Secantes

O Método das Secantes é semelhante ao Método de Newton. Neste método a derivada $f'(x)$ é aproximada pela declividade de um reta secante à curva:

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Assim, em cada passo do método, calcula-se uma nova aproximação com base em duas aproximações anteriores:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{m}, \qquad m = \frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}$$

Exemplo 32. *Encontre as raízes de* $f(x) = \cos(x) - x$.

Da inspeção do gráfico das funções $y = \cos(x)$ e $y = x$, sabemos que esta equação possui uma raiz em torno de x=0,8. Iniciamos o método com $x_0 = 0,7$ e $x_1 = 0,8$.

$x^{(n-1)}$	$x^{(n)}$	m	$x^{(n+1)}$
0,7	0,8	$\frac{f(0,8)-f(0,7)}{0,8-0,7} = -1,6813548$	$0,8 - \frac{f(0,8)}{-1,6813548} = 0,7385654$
$0,8$	$0,7385654$	$-1,6955107$	$0,7390784$
$0,7385654$	$0,7390784$	$-1,6734174$	$0,7390851$
$0,7390784$	$0,7390851$	$-1,6736095$	$0,7390851$

Problema 6. *Aplique o método das secantes para resolver a equação*

$$e^{-x^2} = 2x$$

Problema 7. *Aplique o método das secantes para encontrar as três primeira raízes da função do problema 23 da lista.*

Problema 8. *Resolva novamente o problema do exemplo dado com* $x^{(0)} = 0,8$ *e* $x^{(1)} = 0,7$.

3.5.1 Análise de convergência

Seja $f(x) \in C^2$ um função tal que $f(x^*) = 0$ e $f'(x^*) \neq 0$. Considere o processo iterativo do método das secantes:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})}(x^{(n)} - x^{(n-1)})$$

Esta expressão pode ser escrita como:

$$\begin{aligned} x^{(n+1)} &= x^{(n)} - \frac{f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} \\ &= \frac{x^{(n)}\left(f(x^{(n)}) - f(x^{(n-1)})\right) - f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} \\ &= \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} \end{aligned}$$

Subtraindo x^* de ambos os lados temos:

$$\begin{aligned} x^{(n+1)} - x^* &= \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} - x^* \\ &= \frac{x^{(n)}f(x^{(n-1)}) - x^{(n-1)}f(x^{(n)}) - x^*\left(f(x^{(n)}) - f(x^{(n-1)})\right)}{f(x^{(n)}) - f(x^{(n-1)})} \\ &= \frac{(x^{(n)} - x^*)f(x^{(n-1)}) - (x^{(n-1)} - x^*)f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} \end{aligned}$$

Definimos $\epsilon_n = x_n - x^*$, equivalente a $x_n = x^* + \epsilon_n$

$$\epsilon_{n+1} = \frac{\epsilon_n f(x^* + \epsilon_{n-1}) - \epsilon_{n-1}f(x^* + \epsilon_n)}{f(x^* + \epsilon_n) - f(x^* + \epsilon_{n-1})}$$

Aproximamos a função $f(x)$ no numerador por

$$\begin{aligned} f(x^* + \epsilon) &\approx f(x^*) + \epsilon f'(x^*) + \epsilon^2 \frac{f''(x^*)}{2} \\ f(x^* + \epsilon) &\approx \epsilon f'(x^*) + \epsilon^2 \frac{f''(x^*)}{2} \end{aligned}$$

$$\begin{aligned}\epsilon_{n+1} &\approx \frac{\epsilon_n\left[\epsilon_{n-1}f'(x^*)+\epsilon_{n-1}^2\frac{f''(x^*)}{2}\right]-\epsilon_{n-1}\left[\epsilon_n f'(x^*)+\epsilon_n^2\frac{f''(x^*)}{2}\right]}{f(x^*+\epsilon_n)-f(x^*+\epsilon_{n-1})}\\ &= \frac{\frac{f''(x^*)}{2}\left(\epsilon_n\epsilon_{n-1}^2-\epsilon_{n-1}\epsilon_n^2\right)}{f(x^*+\epsilon_n)-f(x^*+\epsilon_{n-1})}\\ &= \frac{1}{2}\frac{f''(x^*)\epsilon_n\epsilon_{n-1}\left(\epsilon_{n-1}-\epsilon_n\right)}{f(x^*+\epsilon_n)-f(x^*+\epsilon_{n-1})}\end{aligned}$$

Observamos, agora, que

$$\begin{aligned}f(x^*+\epsilon_n)-f(x^*+\epsilon_{n-1}) &\approx [f(x^*)+f'(x^*)\epsilon_n]-[f(x^*)+f'(x^*)\epsilon_{n-1}]\\ &= f'(x^*)(\epsilon_n-\epsilon_{n-1})\end{aligned} \tag{3.1}$$

Portanto:

$$\epsilon_{n+1} \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\epsilon_n\epsilon_{n-1} \tag{3.2}$$

ou, equivalentemente:

$$x^{(n+1)}-x^* \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\left(x^{(n)}-x^*\right)\left(x^{(n-1)}-x^*\right) \tag{3.3}$$

Pode-se mostrar que

$$|x^{(n+1)}-x^*| \approx M|x^{(n)}-x^*|^{\phi}, \quad n \text{ grande} \tag{3.4}$$

com $\phi=\frac{\sqrt{5}+1}{2}\approx 1,618$ e M é uma constante.

Observação 14. *O erro na tabela sempre se refere ao erro absoluto esperado. Nos três últimos métodos, é comum que se exija como critério de parada que a condição seja satisfeita por alguns poucos passos consecutivos. Outros critérios podem ser usados. No métodos das secantes, deve-se ter o cuidado de evitar divisões por zero quando $x_{n+1}-x_n$ muito pequeno em relação à resolução do sistema de numeração.*

Tabela 3.1: Quadro comparativo.

Método	Convergência	Erro	Critério de parada
Bisseção	Linear $(p = 1)$	$\epsilon_{n+1} = \frac{1}{2}\epsilon$	$\frac{b_n - a_n}{2} <$ erro
Iteração linear	Linear $(p = 1)$	$\epsilon_{n+1} \approx \lvert\phi'(x^*)\rvert\varepsilon_n$	$\frac{\lvert\Delta_n\rvert}{1-\frac{\Delta_n}{\Delta_{n-1}}} <$ erro $\Delta_n < \Delta_{n-1}$
Newton	Quadrática $(p = 2)$	$\epsilon_{n+1} \approx \frac{1}{2}\left\lvert\frac{f''(x^*)}{f'(x^*)}\right\rvert\varepsilon_n^2$	$\lvert\Delta_n\rvert <$ erro
Secante	$p = \frac{\sqrt{5}+1}{2}$ $\approx 1,618$	$\varepsilon_{n+1} \approx \left\lvert\frac{f''(x^*)}{f'(x^*)}\right\rvert\varepsilon_n\varepsilon_{n-1}$ $\approx M\varepsilon_n^{\phi}$	$\lvert\Delta_n\rvert <$ erro

Capítulo 4

Solução de sistemas lineares

4.1 Problemas lineares

Neste parte de nosso curso, estamos interessados em técnicas para resolução de sistemas de equações algébricas lineares. O leitor já tem ampla experiência com tais problemas desde o ensino fundamental até o curso de álgebra linear, dedicado à formalização e ao estudo sistematizado de problemas 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 que envolvia no sistema linear é quadrada.

4.2 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 33. Eliminação Gaussiana sem 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:

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

Encontramos $-z=-1$*, ou seja,* $z=1$*. Substituímos na segunda equação e temos* $-y-3z=-2$*, ou seja,* $y=-1$ *e, finalmente* $x+y+z=1$*, resultando em* $x=1$*.*

A Eliminação Gaussiana com pivotamento parcial consiste em fazer uma permutação de linhas de forma a escolher o maior pivô (em módulo) a cada passo.

Exemplo 34. Eliminação Gaussiana com pivotamento parcial*: Resolva o sistema:*

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

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 35.

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

Construímos a matriz completa:

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

Portanto $x = 2$*,* $y = 3$ *e* $z = 1$*.*

Exemplo 36. Problema com elementos com grande diferença de escala

$$\left[\begin{array}{cc} \varepsilon & 2 \\ 1 & \varepsilon \end{array}\right] \left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} 4 \\ 3 \end{array}\right]$$

Executamos a eliminação gaussiana sem pivotamento parcial para $\varepsilon \neq 0$ *e* $|\varepsilon| << 1$*:*

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

Temos

$$y = \frac{3 - 4/\varepsilon}{\varepsilon - 2/\varepsilon}$$

e

$$x = \frac{4 - 2y}{\varepsilon}$$

Observe que a expressão obtida para y se aproximada de 2 quando ε é pequeno:

$$y = \frac{3 - 4/\varepsilon}{\varepsilon - 2/\varepsilon} = \frac{3\varepsilon - 4}{\varepsilon^2 - 2} \longrightarrow \frac{-4}{-2} = 2, \quad \textit{quando } \varepsilon \to 0.$$

Já expressão obtida para x depende justamente da diferença $2 - y$:

$$x = \frac{4 - 2y}{\varepsilon} = \frac{2}{\varepsilon}(2 - y)$$

Assim, quando ε é pequeno, a primeira expressão, implementado em um sistema de ponto flutuante de acurácia finita, produz $y = 2$ e, consequentemente, a expressão para x produz $x = 0$. Isto é, estamos diante um problema de cancelamento catastrófico.

Agora, quando usamos a Eliminação Gaussiana com pivotamento parcial, fazemos uma permutação de linhas de forma a escolher o maior pivô a cada passo:

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

Continuando o procedimento, temos:

$$y = \frac{4 - 4\varepsilon}{2 - \varepsilon^2}$$

e

$$x = 3 - \varepsilon y$$

Observe que tais expressões são analiticamente idênticas às anteriores, no entanto, são mais estáveis numericamente. Quando ε converge a zero, y converge a 2, como no caso anterior. No entanto, mesmo que $y = 2$, a segunda expressão produz $x = 3 - \varepsilon y$, isto é, a aproximação $x \approx 3$ não depende mais de obter $2 - y$ com precisão.

Problema 9. *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 pivotamento parcial (não use o computador para resolver essa questão).

Problema 10. *Calcule a inversa da matriz*

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

usando eliminação Gaussiana com pivotamento parcial.

4.3 Condicionamento de sistemas lineares

4.3.1 Motivação

Quando lidamos com matrizes no corpo do números reais (ou complexos), existem apenas duas alternativas: i) a matriz é inversível; ii) a matriz não é inversível e, neste caso, é chamada de matriz singular. Ao lidarmos em aritmética de precisão finita, encontramos uma situação mais sutil: alguns problema lineares são mais difíceis de serem resolvidos, pois os erros de arredondamento se propagam de forma mais significativa que em outros problemas. Neste caso falamos de problemas bem-condicionados e mal-condicionados. Intuitivamente falando, um problema bem-condicionado é um problema em que os erros de arredondamento se propagam de forma menos importante; enquanto problemas mal-condicionados são problemas em que os erros se propagam de forma mais relevante.

Um caso típico de sistema mal-condicionado é aquele cujos coeficiente estão muito próximos ao de um problema singular. Considere o seguinte exemplo:

Exemplo 37. *Observe que o problema*

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

é impossível quando $\lambda = \frac{71\times 30}{41} \approx 51,95122$.

Agora, verifique o que acontece quando resolvemos os seguintes sistemas lineares:

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

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

Igualmente, observe os seguintes dois problemas:

$$\begin{cases} 71x + 41y = 100 \\ 52x + 30y = 70 \end{cases} \quad e \quad \begin{cases} 71x + 41y = 100,4 \\ 52x + 30y = 69,3 \end{cases}$$

A solução do primeiro problema é $x = -65$ e $y = 115$ e do segundo problema é $x = -85,35$ e $y = 150,25$.

Observe que pequenas variações nos coeficientes das matrizes fazem as soluções ficarem bem distintas, isto é, pequenas variações nos dados de entrada acarretaram em grandes variações na solução do sistema. Quando isso acontece, dizemos que o problema é mal-condicionados.

Para introduzir essa ideia formalmente, precisamos definir o número de condicionamento. Informalmente falando, o número de condicionamento mede o quanto a solução de um problema em função de alterações nos dados de entrada. Para construir matematicamente este conceito, precisamos de uma medida destas variações. Como tanto os dados de entrada como os dados de saída são expressos na forma vetorial, precisaremos do conceito de norma vetorial. Por isso, faremos uma breve interrupção de nossa discussão para introduzir as definições de norma de vetores e matrizes na próxima seção.

4.3.2 Norma L_p de vetores

Definimos a norma L_p ou L^p de um vetor em $\mathbb{R}^n$ para $p \geq 1$ como

$$\|v\|_p = (|v_1|^p + |v_2|^p + \cdots |v_n|^p)^{1/p}$$

E a norma L_∞ ou L^∞ como

$$\|v\|_\infty = \max_{j=1}^{n} |v_j|$$

Propriedades: Se λ é um real (ou complexo) e u e v são vetores, temos:

$$\begin{aligned} \|v\| &= 0 \Longleftrightarrow v = 0 \\ \|\lambda v\| &= |\lambda| \|v\| \\ \|u + v\| &\leq \|u\| + \|v\| \quad \text{(desigualdade do triângulo)} \\ \lim_{p \to \infty} \|u\|_p &= \|u\|_\infty \end{aligned}$$

Exemplo: Calcule a norma L^1, L^2 e L^∞ de

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

$$\begin{aligned} \|v\|_1 &= 1+2+3+0=6 \\ \|v\|_2 &= \sqrt{1+2^2+3^2+0^2}=\sqrt{14} \\ \|v\|_\infty &= \max\{1,2,3,0\}=3 \end{aligned}$$

4.3.3 Norma matricial

Definimos a norma operacional em L^p de uma matriz $A : \mathbb{R}^n \to \mathbb{R}^n$ da seguinte forma:

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

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

Temos as seguintes propriedades, se A e B são matrizes, I é a matriz identidade, v é um vetor e λ é um real (ou complexo):

$$\begin{aligned} \|A\|_p &= 0 \iff A=0 \\ \|\lambda A\|_p &= |\lambda|\|A\|_p \\ \|A+B\|_p &\le \|A\|_p + \|B\|_p \quad \text{(desigualdade do triângulo)} \\ \|Av\|_p &\le \|A\|_p\|v\|_p \\ \|AB\|_p &\le \|A\|_p\|B\|_p \\ \|I\|_p &= 1 \\ 1 &= \|I\|_p = \|AA^{-1}\|_p \le \|A\|_p\|A^{-1}\|_p \quad \text{(se A é inversível)} \end{aligned}$$

Casos especiais:

$$\begin{aligned} \|A\|_1 &= \max_{j=1}^{n} \sum_{i=1}^{n} |A_{ij}| \\ \|A\|_2 &= \sqrt{\max\{|\lambda| : \lambda \in \sigma(AA^*)\}} \\ \|A\|_\infty &= \max_{i=1}^{n} \sum_{j=1}^{n} |A_{ij}| \end{aligned}$$

onde $\sigma(M)$ é o conjunto de autovalores da matriz M.

Exemplo: Calcule as normas 1, 2 e ∞ da seguinte matriz:

$$A = \begin{bmatrix} 3 & -5 & 7 \\ 1 & -2 & 4 \\ -8 & 1 & -7 \end{bmatrix}$$

Solução

$$\begin{aligned} \|A\|_1 &= \max\{12, 8, 18\} = 18 \\ \|A\|_\infty &= \max\{15, 7, 16\} = 16 \\ \|A\|_2 &= \sqrt{\max\{0,5865124; 21,789128; 195,62436\}} = 13,986578 \end{aligned}$$

4.3.4 Número de condicionamento

O condicionamento de um sistema linear é um conceito relacionado à forma como os erros se propagam dos dados de entrada para os dados de saída, ou seja, se o sistema

$$Ax = y$$

possui uma solução x para o vetor y, quando varia a solução x quando o dado de entrado y varia. Consideramos, então, o problema

$$A(x + \delta_x) = y + \delta_y$$

Aqui δ_x representa a variação em x e δ_y representa a respectiva variação em y. Temos:

$$Ax + A\delta_x = y + \delta_y$$

e, portanto,

$$A\delta_x = \delta_y.$$

Queremos avaliar a magnitude do erro relativo em y, representado por $\|\delta_y\|/\|y\|$ em função da magnitude do erro relativo $\|\delta_x\|/\|x\|$.

$$\frac{\|\delta_x\|}{\|x\|} \Big/ \frac{\|\delta_y\|}{\|y\|} = \frac{\|\delta_x\|}{\|x\|}\frac{\|y\|}{\|\delta_y\|} = \frac{\|A^{-1}\delta_y\|}{\|x\|}\frac{\|Ax\|}{\|\delta_y\|} \leq \frac{\|A^{-1}\|\|\delta_y\|}{\|x\|}\frac{\|A\|\|x\|}{\|\delta_y\|} = \|A\|\|A^{-1}\|$$

Assim, definimos o número de condicionamento de uma matriz inversível A como

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

O número de condicionamento, então, mede o quão instável é resolver o problema $Ax = y$ frente a erros no vetor de entrada x.

Obs: O número de condicionamento depende da norma escolhida.

Obs: O número de condicionamento da matriz identidade é 1.

Obs: O número de condicionamento de qualquer matriz inversível é igual ou maior que 1.

Exemplo Calcule o número de condicionamento da matriz

$$A = \begin{bmatrix} 3 & -5 & 7 \\ 1 & -2 & 4 \\ -8 & 1 & -7 \end{bmatrix}$$

nas normas 1, 2 e ∞.
Resp: $k_1(A) = 36$, $k_2(A) = 18,26$, $K_\infty(A) = 20,8$.

4.4 Métodos iterativos para sistemas lineares

4.4.1 Método de Jacobi

Considere o problema $Ax = y$, ou seja,

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & y_2 \\ \vdots \qquad\qquad\qquad\qquad \vdots & = & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & y_n \end{array}$$

Os elementos x_j são calculados iterativamente conforme:

$$\begin{array}{rcl} x_1^{(k+1)} & = & \dfrac{y_1 - \left(a_{12}x_2^{(k)} + \cdots + a_{1n}x_n^{(k)}\right)}{a_{11}} \\ x_2^{(k+1)} & = & \dfrac{y_2 - \left(a_{21}x_1^{(k)} + \cdots + a_{2n}x_n^{(k)}\right)}{a_{22}} \\ & \vdots & \\ x_n^{(k+1)} & = & \dfrac{y_2 - \left(a_{n1}x_1^{(k)} + \cdots + a_{n(n-1)}x_{n-1}^{(k)}\right)}{a_{nn}} \end{array}$$

Em notação mais compacta, o método de Jacobi consiste na iteração:

$$\begin{array}{c} x^{(0)} = \text{aprox. inicial} \\ x_i^{(k)} = \dfrac{y_i - \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)}}{a_{ii}} \end{array}$$

Exemplo: Resolva o sistema

$$\begin{cases} 10x + y = 23 \\ x + 8y = 26 \end{cases}$$

usando o método de Jacobi iniciando com $x^{(0)} = y^{(0)} = 0$.

$$\begin{aligned} x^{(k+1)} &= \frac{23 - y^{(k)}}{10} \\ y^{(k+1)} &= \frac{26 - x^{(k)}}{8} \end{aligned}$$

$$\begin{aligned} x^{(1)} &= \frac{23 - y^{(0)}}{10} = 2,3 \\ y^{(1)} &= \frac{26 - x^{(0)}}{8} = 3,25 \end{aligned}$$

$$\begin{aligned} x^{(2)} &= \frac{23 - y^{(1)}}{10} = 1,975 \\ y^{(2)} &= \frac{26 - x^{(1)}}{8} = 2,9625 \end{aligned}$$

Algoritmo de Jacobi

4.4.2 Método de Gauss-Seidel

Considere o problema $Ax = y$, ou seja,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \vdots \qquad\qquad\qquad\quad \vdots &= \vdots \\ a_{n1}x_1 + a_{22}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned}$$

Os elementos x_j são calculados iterativamente conforme:

$$\begin{aligned} x_1^{(k+1)} &= \frac{y_1 - \left(a_{12}x_2^{(k)} + \cdots + a_{1n}x_n^{(k)}\right)}{a_{11}} \\ x_2^{(k+1)} &= \frac{y_2 - \left(a_{11}x_1^{(k+1)} + \cdots + a_{1n}x_n^{(k)}\right)}{a_{22}} \\ &\vdots \\ x_n^{(k+1)} &= \frac{y_2 - \left(a_{n1}x_1^{(k+1)} + \cdots + a_{n(n-1)}x_{n-1}^{(k+1)}\right)}{A_{nn}} \end{aligned}$$

Em notação mais compacta, o método de Gauss-Seidel consiste na iteração:

$$x^{(0)} = \text{aprox. inicial}$$
$$x_i^{(k)} = \frac{y_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}}{a_{ii}}$$

Exemplo: Resolva o sistema

$$\begin{cases} 10x + y = 23 \\ x + 8y = 26 \end{cases}$$

usando o método de Guass-Seidel iniciando com $x^{(0)} = y^{(0)} = 0$.

$$\begin{aligned} x^{(k+1)} &= \frac{23 - y^{(k)}}{10} \\ y^{(k+1)} &= \frac{26 - x^{(k+1)}}{8} \end{aligned}$$

$$\begin{aligned} x^{(1)} &= \frac{23 - y^{(0)}}{10} = 2,3 \\ y^{(1)} &= \frac{26 - x^{(1)}}{8} = 2,9625 \end{aligned}$$

$$\begin{aligned} x^{(2)} &= \frac{23 - y^{(1)}}{10} = 2,00375 \\ y^{(2)} &= \frac{26 - x^{(2)}}{8} = 2,9995312 \end{aligned}$$

Algoritmo de Gauss-Seidel

4.5 Análise de convergência

Uma condição suficiente porém não necessária para que os métodos de Gauss-Seidel e Jacobi convirjam é a que a matriz seja diagonal dominante estrita. Ver Burden & Faires.

Problema 11. *Resolva o seguinte sistema pelo método de Jacobi e Gauss-Seidel:*

$$\begin{cases} 5x_1 + x_2 + x_3 &= 50 \\ -x_1 + 3x_2 - x_3 &= 10 \\ x_1 + 2x_2 + 10x_3 &= -30 \end{cases}$$

Use como critério de paragem tolerância inferior a 10^{-3} *e inicialize com* $x^0 = y^0 = z^0 = 0$.

4.6 Método da potência para cálculo de autovalores

Consideremos uma matriz $A \in \mathbb{R}^{n,n}$ diagonalizável, isto é, existe um conjunto $\{v_j\}_{j=1}^n$ de autovetores de A tais que qualquer elemento $x \in \mathbb{R}^n$ pode ser escrito como uma combinação linear dos v_j. Sejam $\{\lambda_j\}_{j=1}^n$ o conjunto de autovalores associados aos autovetores tal que um deles seja dominante, ou seja,

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

Como os autovetores são LI, todo vetor $x \in \mathbb{R}^n$, $x = (x_1, x_2, ..., x_n)$, pode ser escrito com combinação linear dos autovetores da seguinte forma:

$$x = \sum_{j=1}^{n} \beta_j v_j. \tag{4.1}$$

O método da potência permite o cálculo do autovetor dominante com base no comportamento assintótico (i.e. "no infinito") da sequência

$$x, Ax, A^2x, A^3x, \ldots$$

.

Por questões de convergência, consideramos a seguinte sequência semelhante à anterior, porém normalizada:

$$\frac{x}{\|x\|}, \frac{Ax}{\|Ax\|}, \frac{A^2x}{\|A^2x\|}, \frac{A^3x}{\|A^3x\|}, \ldots,$$

que pode ser obtida pelo seguinte processo iterativo:

$$x^{(k+1)} = \frac{A^k x}{\|A^k x\|}$$

Observamos que se x está na forma (4.1), então $A^k x$ pode ser escrito como

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

Como $\left|\frac{\lambda_j}{\lambda_1}\right| < 1$ para todo $j \geq 2$, temos

$$\sum_{j=2}^{n} \frac{\beta_j}{\beta_1} \left(\frac{\lambda_j}{\lambda_1}\right)^k v_j \to 0.$$

Assim

$$\frac{A^k x}{\|A^k x\|} = \frac{\beta_1 \lambda_1^k}{\|A^k x\|} \left(v_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right) \tag{4.2}$$

Como a norma de $\frac{A^k x}{\|A^k x\|}$ é igual a um, temos

$$\left\| \frac{\beta_1 \lambda_1^k}{\|A^k x\|} v_1 \right\| \to 1$$

e, portanto,

$$\left| \frac{\beta_1 \lambda_1^k}{\|A^k x\|} \right| \to \frac{1}{\|v_1\|}$$

Ou seja, se definimos $\alpha^{(k)} = \frac{\beta_1 \lambda_1^k}{\|A^k x\|}$, então

$$|\alpha^{(k)}| \to 1$$

Retornando a (4.2), temos:

$$\frac{A^k x}{\|A^k x\|} - \alpha^{(k)} v_1 \to 0$$

Observe que um múltiplo de autovetor também é um autovetor e, portanto,

$$\frac{A^k x}{\|A^k x\|}$$

é um esquema que oscila entre os autovetores ou converge para o autovetor v_1.

Uma vez que temos o autovetor v_1 de A, podemos calcular λ_1 da seguinte forma:

$$Av_1 = \lambda_1 v_1 \implies v_1^T A v_1 = v_1^T \lambda_1 v_1 \implies \lambda_1 = \frac{v_1^T A v_1}{v_1^T v_1}$$

Observe que a última identidade é válida, pois $\|v_1\| = 1$ por construção.

Exercício 17. *Calcule o autovalor dominante e o autovetor associado da matriz*

$$\begin{bmatrix} 3 & 4 \\ 2 & -1 \end{bmatrix}$$

usando o método da potência inciando com o vetor $x = [1 \quad 1]^T$

Exercício 18. *Os autovalores de uma matriz triangular são os elementos da diagonal principal. Verifique o método da potência aplicada à seguinte matriz:*

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

Capítulo 5

Solução de sistemas de equações não lineares

O método de Newton aplicado a encontrar a raiz x^* da função $y = f(x)$ estudado na primeira área de nossa disciplina consiste em um processo iterativo. Em cada passo deste processo, dispomos de uma aproximação $x^{(k)}$ para x^* e construímos uma aproximação $x^{(k+1)}$. Cada passo do método de Newton envolve os seguintes procedimentos:

- Linearização da função $f(x)$ no ponto $x^{(k)}$: $f(x) = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)}) + O\left(|x - x^{(k)}|^2\right)$

- A aproximação $x^{(k+1)}$ é definida como o valor de x em que a linearização $f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)})$ passa por zero.

Observação: $y = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)})$ é a equação da reta que tangencia a curva $y = f(x)$ no ponto $(x^{(k)}, f(x^{(k)}))$.

Queremos, agora, generalizar o método de Newton a fim de resolver problemas de várias equações e várias incógnitas, ou seja, encontrar $x_1, x_2, \ldots x_n$ que satisfazem as seguinte equações:

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

Podemos escrever este problema na forma vetorial definindo o vetor $x = [x_1, x_2, \ldots, x_n]^T$ e a função vetorial

$$F(x) = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) \end{bmatrix}$$

Exemplo 38. *Suponha que queiramos resolver numericamente os seguinte sistema de duas equações e duas incógnitas:*

$$\begin{aligned} \frac{x_1^2}{3} + x_2^2 &= 1 \\ x_1^2 + \frac{x_2^2}{4} &= 1 \end{aligned}$$

Então definimos

$$F(x) = \begin{bmatrix} \frac{x_1^2}{3} + x_2^2 - 1 \\ \\ x_1^2 + \frac{x_2^2}{4} - 1 \end{bmatrix}$$

Neste momento, dispomos de um problema na forma $F(x) = 0$ e precisamos desenvolver uma técnica para linearizar a função $F(x)$. Para tal, precisamos de alguns conceitos do Cálculo II.

Observe que $F(x) - F(x^{(0)})$ pode ser escrito como

$$F(x) - F(x^{(0)}) = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) - f_1(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) \\ f_2(x_1, x_2, \ldots, x_n) - f_2(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) - f_n(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) \end{bmatrix}$$

Usamos a regra da cadeia

$$df_i = \frac{\partial f_i}{\partial x_1} dx_1 + \frac{\partial f_i}{\partial x_2} dx_2 + \cdots + \frac{\partial f_i}{\partial x_n} dx_n = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} dx_j$$

e aproximamos as diferenças por derivadas parciais:

$$f_i(x_1, x_2, \ldots, x_n) - f_i(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) \approx \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \left(x_j - x_j^{(0)} \right)$$

Portanto,

$$F(x) - F(x^{(0)}) \approx \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \begin{bmatrix} x_1 - x_1^{(0)} \\ x_2 - x_2^{(0)} \\ \vdots \\ x_n - x_n^{(0)} \end{bmatrix} \tag{5.1}$$

Definimos então a matriz jacobiana por

$$J_F = \frac{\partial(f_1, f_2, \ldots, f_n)}{\partial(x_1, x_2, \ldots, x_n)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

A matriz jacobiana de uma função ou simplesmente, o Jacobiano de uma função $F(x)$ é a matriz formada pelas suas derivadas parciais:

$$(J_F)_{ij} = \frac{\partial f_i}{\partial x_j}$$

Nestes termos podemos reescrever (5.1) como

$$F(x) \approx F(x^{(0)}) + J_F(x^{(0)})(x - x^{(0)})$$

Esta expressão é chama de linearização de $F(x)$ no ponto $x^{(0)}$ e generaliza a linearização em uma dimensão dada por $f(x) \approx f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$

5.1 O método de Newton para sistemas

Vamos agora construir o método de Newton-Raphson, ou seja, o método de Newton generalizado para sistemas. Assumimos, portanto, que a função $F(x)$ é diferenciável e que existe um ponto x^* tal que $F(x^*) = 0$. Seja $x^{(k)}$ uma aproximação para x^*, queremos construir uma nova aproximação $x^{(k+1)}$ através da linearização de $F(x)$ no ponto $x^{(k)}$.

- Linearização da função $F(x)$ no ponto $x^{(k)}$: $F(x) = F(x^{(k)}) + J_F\left(x^{(k)}\right)\left(x - x^{(k)}\right) + O\left(\|x - x^{(k)}\|^2\right)$

- A aproximação $x^{(k)}$ é definida como o ponto x em que a linearização $F(x^{(k)}) + J_F\left(x^{(k)}\right)\left(x - x^{(k)}\right)$ é nula, ou seja:

$$F(x^{(k)}) + J_F\left(x^{(k)}\right)\left(x^{(k+1)} - x^{(k)}\right) = 0$$

Supondo que a matriz jacobina seja inversível no ponto $x^{(k)}$, temos:

$$\begin{aligned} J_F\left(x^{(k)}\right)\left(x^{(k+1)} - x^{(k)}\right) &= -F(x^{(k)}) \\ x^{(k+1)} - x^{(k)} &= -J_F^{-1}\left(x^{(k)}\right) F(x^{(k)}) \\ x^{(k+1)} &= x^{(k)} - J_F^{-1}\left(x^{(k)}\right) F(x^{(k)}) \end{aligned}$$

Desta forma, o método iterativo de Newton-Raphson para encontrar as raízes de $F(x) = 0$ é dado por

$$\begin{cases} x^{(k+1)} = x^{(k)} - J_F^{-1}\left(x^{(k)}\right) F(x^{(k)}), \quad n \geq 0 \\ \\ x^{(0)} = \text{dado inicial} \end{cases}$$

Observação 15. *Usamos subíndices para indicar o elemento de um vetor e super-índices para indicar o passo da iteração. Assim $x^{(k)}$ se refere à iteração k e $x_i^{(k)}$ se refere à componente i no vetor $x^{(k)}$.*

Observação 16. *A notação $J_F^{-1}\left(x^{(k)}\right)$ enfatiza que a jacobiana deve ser calculada a cada passo.*

Observação 17. *Podemos definir o passo $\Delta^{(k)}$ como*

$$\Delta^{(k)} = x^{(k+1)} - x^{(k)}$$

Assim, $\Delta^{(k)} = -J_F^{-1}\left(x^{(k)}\right) F(x^{(k)})$, ou seja, $\Delta^{(k)}$ resolve o problema linear:

$$J_F\left(x^{(k)}\right)\Delta^{(k)} = -F(x^{(k)})$$

Em geral, é menos custoso resolver o sistema acima do que calcular o inverso da jacobiana e multiplicar pelo vetor $F(x^{(k)})$.

Exemplo 39. *Retornamos ao nosso exemplo inicial, isto é, resolver numericamente os seguinte sistema não-linear:*

$$\begin{aligned} \frac{x_1^2}{3} + x_2^2 &= 1 \\ x_1^2 + \frac{x_2^2}{4} &= 1 \end{aligned}$$

Para tal, definimos a função $F(x)$:

$$F(x) = \begin{bmatrix} \frac{x_1^2}{3} + x_2^2 - 1 \\ \\ x_1^2 + \frac{x_2^2}{4} - 1 \end{bmatrix}$$

cuja jacobiana é

$$J_F = \begin{bmatrix} \frac{2x_1}{3} & 2x_2 \\ \\ 2x_1 & \frac{x_2}{2} \end{bmatrix}$$

Faremos a implementação numérica no Scilab. Para tal definimos as funções que implementarão $F(x)$ e a $J_F(x)$

```
function y=F(x)
    y(1)=x(1)^2/3+x(2)^2-1
    y(2)=x(1)^2+x(2)^2/4-1
endfunction

function y=JF(x)
    y(1,1)=2*x(1)/3
    y(1,2)=2*x(2)
    y(2,1)=2*x(1)
    y(2,2)=x(2)/2
endfunction
```

Alternativamente, estas funções poderiam ser escritas como

```
function y=F(x)
    y=[x(1)^2/3+x(2)^2-1; x(1)^2+x(2)^2/4-1]
endfunction

function y=JF(x)
    y=[2*x(1)/3  2*x(2); 2*x(1) x(2)/2]
endfunction
```

Desta forma, se x é uma aproximação para a raiz, pode-se calcular a próxima aproximação através dos comandos:

```
delta=-JF(x)\F(x)
x=x+delta
```

Ou simplesmente

```
x=x-JF(x)\F(x)
```

Observe que as soluções exatas desse sistema são $\left(\pm\sqrt{\frac{9}{11}}, \pm\sqrt{\frac{8}{11}}\right)$

Problema 12. *Encontre uma aproximação para a solução do sistema*

$$\begin{aligned} x_1^2 &= \cos(x_1 x_2) + 1 \\ \sin(x_2) &= 2\cos(x_1) \end{aligned}$$

que fica próxima ao ponto $x_1 = 1.5$ *e* $x_2 = .5$. **Resp:** *(1,3468109, 0,4603195)*

Solução: Definimos a função $F(x)$ como

$$F(x) = \begin{bmatrix} x_1^2 - \cos(x_1 x_2) - 1 \\ \\ \sin(x_2) - 2\cos(x_1) \end{bmatrix}$$

cuja jacobiana é

$$J_F(x) = \begin{bmatrix} 2x_1 + x_2\sin(x_1 x_2) & x_1\sin(x_1 x_2) \\ \\ 2\sin(x_1) & \cos(x_2) \end{bmatrix}$$

Implementamos no Scilab como

```
function y=F(x)
    y=[x(1)^2-cos(x(1)*x(2))-1 ; sin(x(2))-2*cos(x(1))]
endfunction

function y=JF(x)
    y=[2*x(1)+x(2)*sin(x(1)*x(2)) x(1)*sin(x(1)*x(2));2*sin(x(1)) cos(x(2))]
endfunction
```

E agora, basta iterar:

```
x=[1.5; .5]
x=x-JF(x)\F(x) (5 vezes)
```

Problema 13. *Encontre uma aproximação numérica para o seguinte problema não-linear de três equações e três incógnitas:*

$$\begin{aligned} 2x_1 - x_2 &= \cos(x_1) \\ -x_1 + 2x_2 - x_3 &= \cos(x_2) \\ -x_2 + x_3 &= \cos(x_3) \end{aligned}$$

Partindo das seguintes aproximações iniciais:

a) $x^{(0)} = [1,\ 1,\ 1]^T$

b) $x^{(0)} = [-0,5,\ -2,\ -3]^T$

c) $x^{(0)} = [-2,\ -3,\ -4]^T$

d) $x^{(0)} = [0,\ 0,\ 0]^T$

Implementação no scilab:

```
function y=F(x)
y=[
   2*x(1)-x(2)-cos(x(1)) ;
   -x(1)+2*x(2)-x(3)-cos(x(2));
   -x(2)+x(3)-cos(x(3))]
endfunction

function y=JF(x)
y=[
   2+sin(x(1)) -1 0;
   -1 2+sin(x(2)) -1;
   0 -1 1+sin(x(3))]

endfunction
```

5.1.1 Algoritmo de Newton para Sistemas

```
function [x] = newton(F,JF,x0,TOL,N)
  x = x0
  k = 1
  //iteracoes
  while (k <= N)
    //iteracao de Newton
    delta = -inv(JF(x))*F(x)
```

```
    x = x + delta
    //criterio de parada
    if (norm(delta,'inf')<TOL) then
      return x
    end
    k = k+1
  end
  error('Num. de iter. max. atingido!')
endfunction
```

5.2 Linearização de uma função de várias variáveis, o gradiente e a Jacobiana*

5.2.1 O gradiente

Considere primeiramente uma função $f : \mathbb{R}^n \to \mathbb{R}$, ou seja, uma função que mapeia n variáveis reais em um único real, por exemplo:

$$f(x) = x_1^2 + x_2^2/4$$

Para construirmos a linearização, fixemos uma direção no espaço $\mathbb{R}^n$, ou seja um vetor v:

$$v = [v_1, v_2, \cdots, v_n]^T$$

Queremos estudar como a função $f(x)$ varia quando "andamos" na direção v a partir do ponto $x^{(0)}$. Para tal, inserimos um parâmetro real pequeno h, dizemos que

$$x = x^{(0)} + hv$$

e definimos a função auxiliar

$$g(h) = f(x^0 + hv).$$

Observamos que a função $g(h)$ é uma função de $\mathbb{R}$ em $\mathbb{R}$.

A linearização de $g(h)$ em torno de $h = 0$ é dada por

$$g(h) = g(0) + hg'(0) + O(h^2)$$

Observamos que $g(h) = f(x^{(0)} + hv)$ e $g(0) = f(x^{(0)})$. Precisamos calcular $g'(0)$:

$$g'(h) = \frac{d}{dh}g(h) = \frac{d}{dh}f(x^{(0)} + hv)$$

Pela regra da cadeia temos:

$$\frac{d}{dh}f(x^{(0)}+hv)=\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}\frac{dx_j}{dh}$$

Observamos que $x_j = x_j^{(0)} + hv_j$, portanto

$$\frac{dx_j}{dh}=v_j$$

Assim:

$$\frac{d}{dh}f(x^{(0)}+hv)=\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}v_j$$

Observamos que esta expressão pode ser vista como o produto interno entre o gradiente de f e o vetor v:

$$\nabla f=\begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \\ \frac{\partial f}{\partial x_2}\\ \\ \vdots\\ \\ \frac{\partial f}{\partial x_n}\end{bmatrix}\qquad v=\begin{bmatrix}v_1\\ \\ v_2\\ \\ \vdots\\ \\ v_n\end{bmatrix}$$

Na notação cálculo vetorial escrevemos este produto interno como $\nabla f \cdot v = v \cdot \nabla f$ na notação de produto matricial, escrevemos $(\nabla f)^T v = v^T \nabla f$. Esta quantidade é conhecida como **derivada direcional** de f no ponto $x^{(0)}$ na direção v, sobretudo quando $\|v\| = 1$.

Podemos escrever a linearização $g(h) = g(0) + hg'(0) + O(h^2)$ como

$$f(x^{(0)}+hv)=f(x^{(0)})+h\nabla^T f(x^{(0)})\,v+O(h^2)$$

Finalmente, escrevemos $x = x^{(0)} + hv$, ou seja, $hv = x - x^{(0)}$

$$f(x)=f(x^{(0)})+\nabla^T f(x^{(0)})\,(x-x^{(0)})+O(\|x-x^{(0)}\|^2)$$

Observação 18. *Observe a semelhança com a linearização no caso em uma dimensão. A notação $\nabla^T f(x^{(0)})$ é o transposto do vetor gradiente associado à função $f(x)$ no ponto $x^{(0)}$:*

$$\nabla^T f(x^{(0)})=\left[\frac{\partial f\left(x^{(0)}\right)}{\partial x_1},\quad \frac{\partial f\left(x^{(0)}\right)}{\partial x_2},\quad \cdots,\quad \frac{\partial f\left(x^{(0)}\right)}{\partial x_n}\right]$$

5.2.2 A matriz jacobiana

Interessamo-nos, agora, pela linearização da função $F : \mathbb{R}^n \to \mathbb{R}^n$. Lembramos que $F(x)$ pode ser escrita como um vetor de funções $f_j : \mathbb{R}^{\times} \to \mathbb{R}$:

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

Linearizando cada uma das funções f_j, temos:

$$F(x) = \underbrace{\begin{bmatrix} f_1\left(x^{(0)}\right) + \nabla^T f_1(x^{(0)})\ \left(x - x^{(0)}\right) + O(\|x - x^{(0)}\|^2) \\ f_2\left(x^{(0)}\right) + \nabla^T f_2(x^{(0)})\ \left(x - x^{(0)}\right) + O(\|x - x^{(0)}\|^2) \\ \vdots \\ f_n\left(x^{(0)}\right) + \nabla^T f_n(x^{(0)})\ \left(x - x^{(0)}\right) + O(\|x - x^{(0)}\|^2) \end{bmatrix}}_{\text{Vetor coluna}}$$

$$= \underbrace{\begin{bmatrix} f_1\left(x^{(0)}\right) \\ f_2\left(x^{(0)}\right) \\ \vdots \\ f_n\left(x^{(0)}\right) \end{bmatrix}}_{\text{Vetor coluna}} + \underbrace{\begin{bmatrix} \nabla^T f_1(x^{(0)}) \\ \nabla^T f_2(x^{(0)}) \\ \vdots \\ \nabla^T f_n(x^{(0)}) \end{bmatrix}}_{\text{Matriz jacobiana}} \underbrace{\left(x - x^{(0)}\right)}_{\text{Vetor coluna}} + O(\|x - x^{(0)}\|^2)$$

Podemos escrever a linearização de $F(x)$ na seguinte forma mais enxuta:

$$F(x) = F\left(x^{(0)}\right) + J_F(x^{(0)})\left(x - x^{(0)}\right) + O\left(\left\|x - x^{(0)}\right\|^2\right)$$

A matriz jacobiana J_F é matriz cujas linhas são os gradientes transpostos de f_j, ou seja:

$$J_F = \frac{\partial(f_1, f_2, \ldots, f_n)}{\partial(x_1, x_2, \ldots, x_n)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

A matriz jacobiana de uma função ou simplesmente, o Jacobiano de uma função $F(x)$ é a matriz formada pelas suas derivadas parciais:

$$(J_F)_{ij} = \frac{\partial f_i}{\partial x_j}$$

Exemplo 40. *Calcule a matriz jacobiana da função*

$$F(x) = \begin{bmatrix} \frac{x_1^2}{3} + x_2^2 - 1 \\ \\ x_1^2 + \frac{x_2^2}{4} - 1 \end{bmatrix}$$

$$J_F = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{2x_1}{3} & 2x_2 \\ \\ 2x_1 & \frac{x_2}{2} \end{bmatrix}$$

Capítulo 6

Aproximação de funções

O problema geral da interpolação pode ser definido da seguinte forma:

Seja $\mathcal{F}$ uma família de funções $f : D \to E$ e $\{(x_i, y_i)\}_{i=1}^{N}$ um conjunto de pares ordenados tais que $x_i \in D$ e $y_i \in E$, encontrar uma função f da família dada tal que $f(x_i) = y_i$ para cada $1 \leq i \leq N$.

Exemplo 41. *Encontrar uma função $f(x)$ da forma $f(x) = ae^{bx}$ onde a e b são constantes tal que $f(1) = 1$ e $f(2) = 5$. Este problema equivale a resolver o seguinte sistema de equações:*

$$\begin{aligned} ae^{b} &= 1 \\ ae^{2b} &= 5 \end{aligned}$$

Dividindo a segunda equação pela primeira, temos $e^b = 5$, logo, $b = \ln(5)$. Substituindo este valor em qualquer das equações, temos $a = \frac{1}{5}$. Assim

$$f(x) = \frac{1}{5}e^{\ln(5)x} = \frac{1}{5}5^x = 5^{x-1}.$$

Exemplo 42. *Encontrar a função polinomial do tipo $f(x) = a+bx+cx^2$ que passe pelos pontos $(-1, 2)$, $(0, 1)$, $(1, 6)$. Observamos que podemos encontrar os coeficientes a, b e c através do seguinte sistema linear:*

$$\begin{aligned} a - b + c &= 2 \\ a &= 1 \\ a + b + c &= 6 \end{aligned}$$

cuja solução é dada por $a = 1$, $b = 2$ e $c = 3$. Portanto

$$f(x) = 1 + 2x + 3x^2.$$

6.1 Interpolação polinomial

Interpolação polinomial é o caso particular do problema geral de interpolação quando a família de funções é constituída de polinômios.

Teorema 3. *Seja* $\{(x_i, y_i)\}_{i=0}^{n}$ *um conjunto de* $n+1$ *pares ordenados de números reais tais que*

$$i \neq j \Longrightarrow x_i \neq x_j \qquad \textit{(i.e. as abscissas são distintas)}$$

então existe um único polinômio $P(x)$ *de grau igual ou inferior a* n *que passa por todos os pontos dados.*

Demonstração. Observamos que o problema de encontrar os coeficientes a_0, a_1,..., a_n do polinômio

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots a_n x^n = \sum_{k=0}^{n} a_k x^k$$

tal que $P(x_i) = y_i$ é equivalente ao seguinte sistema linear de $n+1$ equações e $n+1$ incógnitas:

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_2 + a_2 x_1^2 + \cdots + a_n x_1^n &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n &= y_n \end{aligned}$$

que pode ser escrito na forma matricial como

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A matriz envolvida é uma matriz de Vandermonde de ordem $n+1$ cujo determinante é dado por

$$\prod_{0 \leq i < j \leq n} (x_j - x_i)$$

É fácil ver que se as abscissas são diferentes dois a dois, então o determinante é não-nulo. Disto decorre que o sistema possui um a solução e que esta solução é única. □

Exemplo 43. *Encontre o polinômio da forma* $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ *que passa pelos pontos*

$$(0,1), (1,2), (2,4), (3,8)$$

Este problema é equivalente ao seguinte sistema linear:

$$\begin{aligned} a_0 &= 1 \\ a_0 + a_1 + a_2 + a_3 &= 2 \\ a_0 + 2a_1 + 4a_2 + 8a_3 &= 4 \\ a_0 + 3a_1 + 9a_2 + 27a_3 &= 8 \end{aligned}$$

cuja solução é $a_0 = 1$, $a_1 = \frac{5}{6}$, $a_2 = 0$ *e* $a_3 = \frac{1}{6}$. *Portanto*

$$P(x) = 1 + \frac{5}{6}x + \frac{1}{6}x^3$$

Exemplo 44. *Encontre o polinômio da forma* $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ *que passa pelos pontos*

$$(0,0), (1,1), (2,4), (3,9)$$

Este problema é equivalente ao seguinte sistema linear:

$$\begin{aligned} a_0 &= 0 \\ a_0 + a_1 + a_2 + a_3 &= 1 \\ a_0 + 2a_1 + 4a_2 + 8a_3 &= 4 \\ a_0 + 3a_1 + 9a_2 + 27a_3 &= 9 \end{aligned}$$

cuja solução é $a_0 = 0$, $a_1 = 0$, $a_2 = 1$ *e* $a_3 = 0$. *Portanto*

$$P(x) = x^2$$

Esta abordagem direta que fizemos ao calcular os coeficientes do polinômio na base canônica se mostra ineficiente quando o número de pontos é grande e quando existe grande discrepância nas abscissas. Neste caso a matriz de Vandermonde é mal-condicionada (ver [5]), acarretando um aumento dos erros de arredondamento na solução do sistema.

Uma maneira de resolver este problema é escrever o polinômio em uma base que produza um sistema mais bem-condicionado.

6.2 Diferenças divididas de Newton

O método das diferenças divididas de Newton consistem em construir o polinômio interpolador da seguinte forma:

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

Assim, o problema de calcular os coeficientes a_0, a_1, ..., a_n é equivalente ao seguinte sistema linear:

$$\begin{aligned}
a_0 &= y_0 \\
a_0 + a_1(x_1 - x_0) &= y_1 \\
a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) &= y_2 \\
&\vdots \\
a_0 + a_1(x_n - x_0) + a_2(x_n - x_0)(x_n - x_1) + \cdots + a_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) &= y_n
\end{aligned}$$

Equivalente à sua forma matricial:

$$\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
1 & (x_1 - x_0) & 0 & \cdots & 0 \\
1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & (x_n - x_0) & (x_n - x_0)(x_n - x_1) & \cdots & (x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} =
\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Este é um sistema triangular inferior que pode ser facilmente resolvido conforme:

$$\begin{aligned}
a_0 &= y_0 \\
a_1 &= \frac{y_1 - a_0}{x_1 - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \\
a_2 &= \frac{y_2 - a_1(x_2 - x_0) - a_0}{(x_2 - x_0)(x_2 - x_1)} = \frac{\frac{y_2 - y_1}{(x_2 - x_1)} - \frac{y_1 - y_0}{(x_1 - x_0)}}{(x_2 - x_0)} \\
&\cdots
\end{aligned}$$

A solução deste sistema pode ser escrita em termos das Diferenças Divididas de Newton, definidas recursivamente conforme:

$$\begin{aligned}
f[x_j] &= y_j \\
f[x_j, x_{j+1}] &= \frac{f[x_{j+1}] - f[x_j]}{x_{j+1} - x_j} \\
f[x_j, x_{j+1}, x_{j+2}] &= \frac{f[x_{j+1}, x_{j+2}] - f[x_j, x_{j+1}]}{x_{j+2} - x_j} \\
&\vdots
\end{aligned}$$

Nesta notação, temos $a_k = f[x_0, x_1, x_2, \ldots, x_k]$

Podemos esquematizar o método na seguinte tabela:

j	x_j	$f[x_j]$	$f[x_{j-1}, x_j]$	$f[x_{j-2}, x_{j-1}, x_j]$
0	x_0	$f[x_0]$		
			$f[x_0, x_1] = \frac{f[x_1]-f[x_0]}{x_1-x_0}$	
1	x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$
			$f[x_1, x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1}$	
2	x_2	$f[x_2]$		

Exemplo 45. *Encontrar o polinômio que passe pelos seguintes pontos*

$$(-1, 3), (0, 1), (1, 3), (3, 43)$$

j	x_j	$f[x_j]$	$f[x_{j-1}, x_j]$	$f[x_{j-2}, x_{j-1}, x_j]$	$f[x_{j-3}, x_{j-2}, x_{j-1}, x_j]$
0	-1	3			
			$\frac{1-3}{0-(-1)} = -2$		
1	0	1		$\frac{2-(-2)}{1-(-1)} = 2$	
			$\frac{3-1}{1-0} = 2$		$\frac{6-2}{3-(-1)} = 1$
2	1	3		$\frac{20-2}{3-0} = 6$	
			$\frac{43-3}{3-1} = 20$		
3	3	43			

Portanto

$$\begin{aligned} P(x) &= 3 - 2(x+1) + 2(x+1)x + (x+1)x(x-1) \\ &= x^3 + 2x^2 - x + 1 \end{aligned}$$

Problema 14. *Considere o seguinte conjunto de pontos:*

$$(-2,-47),(0,-3),(1,4)(2,41)$$

. Encontre o polinômio interpolador usando os métodos vistos. Trace os pontos no Scilab usando o comando 'plot2d' e trace o gráfico do polinômio usando comandos de plotagem e a estrutura de polinômio. **Resp:** $5x^3+2x-3$

6.3 Polinômios de Lagrange

Outra maneira clássica de resolver o problema da interpolação polinomial é através do polinômios de Lagrange. Dado um conjunto de pontos $\{x_j\}_{j=1}^n$ distintos dois a dois, definimos os polinômios de Lagrange como os polinômios de grau $n-1$ que satisfazem as seguintes condições:

$$L_k(x_j) = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}$$

Assim, a solução do problema de encontrar os polinômios de grau $n-1$ tais $P(x_j) = y_j, j = 1, \cdots, n$ é dado por

$$P(x) = y_1L_1(x) + y_2L_2(x) + \cdots + y_nL_n(x) = \sum_{j=1}^{n} y_jL_j(x)$$

Para construir os polinômios de Lagrange, basta olhar para sua forma fatorada, ou seja:

$$L_k(x) = C_k \prod_{1\leq j\neq k\leq n} (x - x_j)$$

onde o coeficiente C_k é obtido da condição $L_k(x_k) = 1$:

$$L_k(x_k) = C_k \prod_{1\leq j\neq k\leq n} (x_k - x_j) \Longrightarrow C_k = \frac{1}{\prod_{1\leq j\neq k\leq n}(x_k - x_j)}$$

Portanto,

$$L_k(x) = \prod_{1\leq j\neq k\leq n} \frac{(x - x_j)}{(x_k - x_j)}$$

Observação 19. *O problema de interpolação quando escrito usando como base os polinômios de Lagrange produz um sistema linear diagonal.*

Exemplo 46. *Encontre o polinômio da forma* $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ *que passa pelos pontos*

$$(0,0), (1,1), (2,4), (3,9)$$

Escrevemos:

$$\begin{aligned} L_1(x) &= \frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} = -\frac{1}{6}x^3 + x^2 - \frac{11}{6}x + 1 \\ L_2(x) &= \frac{x(x-2)(x-3)}{1(1-2)(1-3)} = \frac{1}{2}x^3 - \frac{5}{2}x^2 + 3x \\ L_3(x) &= \frac{x(x-1)(x-3)}{2(2-1)(2-3)} = -\frac{1}{2}x^3 + 2x^2 - \frac{3}{2}x \\ L_4(x) &= \frac{x(x-1)(x-2)}{3(3-1)(3-2)} = \frac{1}{6}x^3 - \frac{1}{2}x^2 + \frac{1}{3}x \end{aligned}$$

Assim temos:

$$P(x) = 0 \cdot L_1(x) + 1 \cdot L_2(x) + 4 \cdot L_3(x) + 9 \cdot L_4(x) = x^2$$

Exemplo 47. *Encontre o polinômio da forma* $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ *que passa pelos pontos*

$$(0,0), (1,1), (2,0), (3,1)$$

Como as abscissas são as mesmas do exemplo anterior, podemos utilizar os mesmos polinômios de Lagrange, assim temos:

$$P(x) = 0 \cdot L_1(x) + 1 \cdot L_2(x) + 0 \cdot L_3(x) + 1 \cdot L_4(x) = \frac{2}{3}x^3 - 3x^2 + \frac{10}{3}x$$

6.4 Aproximação de funções reais por polinômios interpoladores

Teorema 4. *Dados* $n+1$ *pontos distintos,* $x_0,\ x_1,\ \cdots,\ x_n$*, dentro de um intervalo* $[a,b]$ *e uma função* f *com* $n+1$ *derivadas contínuas nesse intervalo* ($f \in C^{n+1}[a,b]$)*, então para cada* x *em* $[a,b]$*, existe um número* $\xi(x)$ *em* (a,b) *tal que*

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n),$$

onde $P(x)$ é o polinômio interpolador. Em especial, pode-se dizer que

$$|f(x) - P(x)| \leq \frac{M}{(n+1)!} |(x - x_0)(x - x_1) \cdots (x - x_n)|,$$

onde

$$M = \max_{x \in [a,b]} |f^{(n+1)}(\xi(x))|$$

Exemplo 48. *Considere a função $f(x) = \cos(x)$ e o polinômio $P(x)$ de grau 2 tal que $P(0) = \cos(0) = 1$, $P(\frac{1}{2}) = \cos(\frac{1}{2})$ e $P(1) = \cos(1)$. Use a fórmula de Lagrange para encontrar $P(x)$. Encontre o erro máximo que se assume ao aproximar o valor de $\cos(x)$ pelo de $P(x)$ no intervalo $[0, 1]$. Trace os gráficos de $f(x)$ e $P(x)$ no intervalo $[0, 1]$ no mesmo plano cartesiano e, depois, trace o gráfico da diferença $\cos(x) - P(x)$. Encontre o erro efetivo máximo $|\cos(x) - P(x)|$.*

$$\begin{aligned} P(x) &= 1\frac{(x-\frac{1}{2})(x-1)}{(0-\frac{1}{2})(0-1)} + \cos\left(\frac{1}{2}\right)\frac{(x-0)(x-1)}{(\frac{1}{2}-0)(\frac{1}{2}-1)} + \cos(1)\frac{(x-0)(x-\frac{1}{2})}{(1-0)(1-\frac{1}{2})} \\ &\approx 1 - 0,0299720583066x - 0,4297256358252x^2 \end{aligned}$$

```
L1=poly([.5 1],'x');L1=L1/horner(L1,0)
L2=poly([0 1],'x');L2=L2/horner(L2,0.5)
L3=poly([0 .5],'x');L3=L3/horner(L3,1)
P=L1+cos(.5)*L2+cos(1)*L3
x=[0:.05:1]
plot(x,cos)
plot(x,horner(P,x),'red')
plot(x,horner(P,x)-cos(x))
```

Para encontrar o erro máximo, precisamos estimar $|f'''(x)| = |\sin(x)| \leq \sin(1) < 0,85$ e

$$\max_{x \in [0,1]} \left| x\left(x - \frac{1}{2}\right)(x-1) \right|$$

O polinômio de grau três $Q(x) = x\left(x - \frac{1}{2}\right)(x-1)$ tem um mínimo (negativo) em $x_1 = \frac{3+\sqrt{3}}{6}$ e um máximo (positivo) em $x_2 = \frac{3-\sqrt{3}}{6}$. Logo:

$$\max_{x \in [0,1]} \left| x\left(x - \frac{1}{2}\right)(x-1) \right| \leq \max\{|Q(x_1)|,\ |Q(x_2)|\} \approx 0,0481125.$$

Portanto:

$$|f(x) - P(x)| < \frac{0,85}{3!} 0,0481125 \approx 0,0068159 < 7 \cdot 10^{-3}$$

Para encontrar o erro efetivo máximo, basta encontrar o máximo de $|P(x) - cos(x)|$. O mínimo (negativo) de $P(x) - cos(x)$ acontece em $x_1 = 4,29 \cdot 10^{-3}$ e o máximo (positivo) acontece em $x_2 = 3,29 \cdot 10^{-3}$. Portanto, o erro máximo efetivo é $4,29 \cdot 10^{-3}$.

Exemplo 49. *Considere o problema de aproximar o valor da integral $\int_0^1 f(x)dx$ pelo valor da integral do polinômio $P(x)$ que coincide com $f(x)$ nos pontos $x_0 = 0$, $x_1 = \frac{1}{2}$ e $x_2 = 1$. Use a fórmula de Lagrange para encontrar $P(x)$. Obtenha o valor de $\int_0^1 f(x)dx$ e encontre uma expressão para o erro de truncamento.*

O polinômio interpolador de $f(x)$ é

$$\begin{aligned} P(x) &= f(0)\frac{(x-\frac{1}{2})(x-1)}{(0-\frac{1}{2})(0-1)} + f\left(\frac{1}{2}\right)\frac{(x-0)(x-1)}{(\frac{1}{2}-0)(\frac{1}{2}-1)} + f(1)\frac{(x-0)(x-\frac{1}{2})}{(1-0)(1-\frac{1}{2})} \\ &= f(0)(2x^2-3x+1) + f\left(\frac{1}{2}\right)(-4x^2+4x) + f(1)(2x^2-x) \end{aligned}$$

e a integral de $P(x)$ é

$$\begin{aligned} \int_0^1 P(x)dx &= \left[f(0)\left(\frac{2}{3}x^3 - \frac{3}{2}x^2 + x\right) + f\left(\frac{1}{2}\right)\left(-\frac{4}{3}x^3 + 2x^2\right) + f(1)\left(\frac{2}{3}x^3 - \frac{1}{2}x^2\right)\right]_0^1 \\ &= f(0)\left(\frac{2}{3} - \frac{3}{2} + 1\right) + f\left(\frac{1}{2}\right)\left(-\frac{4}{3} + 2\right) + f(1)\left(\frac{2}{3} - \frac{1}{2}\right) \\ &= \frac{1}{6}f(0) + \frac{2}{3}f\left(\frac{1}{2}\right) + \frac{1}{6}f(1) \end{aligned}$$

Para fazer a estimativa de erro usando o teorema (4), e temos

$$\begin{aligned} \left|\int_0^1 f(x)dx - \int_0^1 P(x)dx\right| &= \left|\int_0^1 f(x) - P(x)dx\right| \\ &\leq \int_0^1 |f(x) - P(x)|dx \\ &\leq \frac{M}{6}\int_0^1 \left|x\left(x-\frac{1}{2}\right)(x-1)\right|dx \\ &= \frac{M}{6}\left[\int_0^{1/2} x\left(x-\frac{1}{2}\right)(x-1)dx - \int_{1/2}^1 x\left(x-\frac{1}{2}\right)(x-1)dx\right] \\ &= \frac{M}{6}\left[\frac{1}{64} - \left(-\frac{1}{64}\right)\right] = \frac{M}{192}. \end{aligned}$$

Lembramos que $M = \max_{x\in[0,1]} |f'''(x)|$.

Observação 20. *Existem estimativas melhores para o erro de truncamento para este esquema de integração numérica. Veremos com mais detalhes tais esquemas na teoria de integração numérica.*

Problema 15. *Use o resultado do exemplo anterior para aproximar o valor das seguintes integrais:*

a) $\int_0^1 \ln(x+1)dx$

b) $\int_0^1 e^{-x^2}dx$

Usando a fórmula obtida, temos que

$$\int_0^1 \ln(x+1)dx \approx 0,39 \pm \frac{1}{96}$$

$$\int_0^1 e^{-x^2}dx \approx 0,75 \pm \frac{3,87}{192}$$

Problema 16. *Use as mesmas técnicas usadas o resultado do exemplo (49) para obter uma aproximação do valor de*

$$\int_0^1 f(x)dx$$

através do polinômio interpolador que coincide com $f(x)$ *nos pontos* $x = 0$ *e* $x = 1$*.*

Resp: $\int_0^1 P(x)dx = \frac{f(0)+f(1)}{2}$, $\frac{1}{12}\max_{x\in[0,1]}|f''(x)|$

6.5 Ajuste de curvas pelo método dos mínimos quadrados

No problema de interpolação, desejamos encontrar uma função $f(x)$ tal que

$$f(x_j) = y_j$$

para um conjunto de pontos dados.

Existem diversas situações em que desejamos encontrar uma função que se aproxime desses pontos.

No problema de ajuste de curvas, busca-se a função $f(x)$ de família de funções dadas que melhor se aproxima de um conjunto de pontos dados. O critério mais usado para o ajuste é critério dos mínimos quadrados, ou seja, buscamos a função $f(x)$ da família que minimiza a soma dos erros elevados ao quadrado:

$$E_q = [f(x_1) - y_1]^2 + [f(x_2) - y_2]^2 + \cdots + [f(x_n) - y_n]^2 = \sum_{j=1}^{n} [f(x_j) - y_j]^2$$

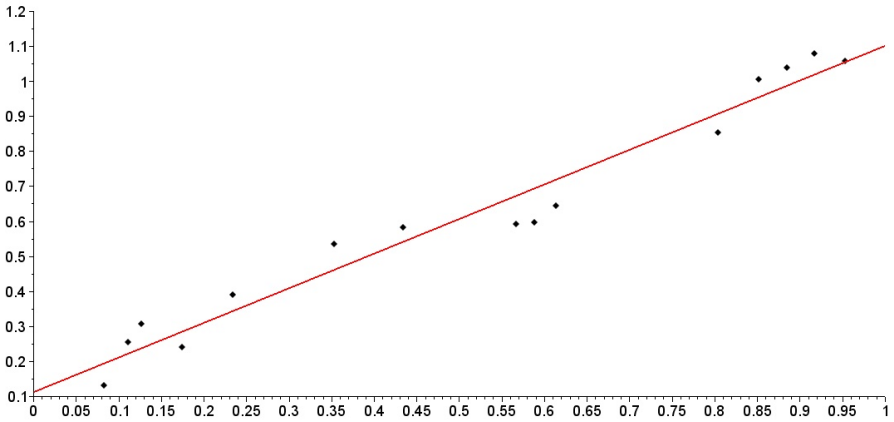


Figura 6.1: Conjunto de 15 pontos e a reta que melhor se ajuste a eles pelo critério do mínimos quadrados.

Exemplo 50. *Encontre a função do tipo* $f(x) = ax$ *que melhor se aproxima dos seguintes pontos:*

$$(0, -0,1), (1,\ 2), (2,\ 3,7)\ e\ (3,\ 7).$$

Defina

$$E_q = [f(x_1) - y_1]^2 + [f(x_2) - y_2]^2 + [f(x_3) - y_3]^2 + [f(x_4) - y_4]^2$$

temos que

$$\begin{aligned} E_q &= [f(0) + 0,1]^2 + [f(1) - 2]^2 + [f(2) - 3,7]^2 + [f(3) - 7]^2 \\ &= [0,1]^2 + [a - 2]^2 + [2a - 3,7]^2 + [3a - 7]^2 \end{aligned}$$

Devemos encontrar o parâmetro a que minimiza o erro, portanto, calculamos:

$$\frac{\partial E_q}{\partial a} = 2[a - 2] + 4[2a - 3,7] + 6[3a - 7] = 28a - 60,8$$

Portanto o valor de a que minimiza o erro é $a = \frac{60,8}{28}$.

```
x=[0 1 2 3]'
y=[-.1 2 3.7 7]'
plot2d(x,y,style=-4)
```

Problema 17. *Encontre a função do tipo $f(x) = bx+a$ que melhor aproxima os pontos do problema anterior.*

Resp: $f(x) = -0,3 + 2,3x$

$$\begin{aligned} E_q &= [f(0) + 0,1]^2 + [f(1) - 2]^2 + [f(2) - 3,7]^2 + [f(3) - 7]^2 \\ &= [a + 0,1]^2 + [a + b - 2]^2 + [a + 2b - 3,7]^2 + [a + 3b - 7]^2 \end{aligned}$$

Devemos encontrar os parâmetros a b que minimizam o erro, por isso, calculamos as derivadas parciais:

$$\begin{aligned} \frac{\partial E_q}{\partial a} &= 2[a + 0,1] + 2[a + b - 2] + 2[a + 2b - 3,7] + 2[a + 3b - 7] \\ \frac{\partial E_q}{\partial b} &= 2[a + b - 2] + 4[a + 2b - 3,7] + 6[a + 3b - 7] \end{aligned}$$

O erro mínimo acontece quando as derivadas são nulas, ou seja:

$$\begin{aligned} 8a + 12b &= 25,2 \\ 12a + 28b &= 60,8 \end{aligned}$$

Cuja solução é dada por $a = -0,3$ e $b = 2,3$. Portanto a função que procuramos é $f(x) = -0,3 + 2,3x$.

6.6 O caso linear

6.6.1 Revisão de Álgebra Linear - O método dos mínimos quadrados para problemas lineares impossíveis

Considere o sistema linear dado por $Ax = b$ onde A é uma matriz $n \times m$ e b é um vetor de n linhas. Assumimos as seguintes hipóteses:

- $n \geq m$. O número de linhas é igual ou superior ao número de colunas. (Mais equações que incógnitas)
- O posto de A é m, i.e., existem m linhas L.I. Isso implica que $Av = 0$ apenas quando $v = 0$

Neste caso, não seremos necessariamente capazes de encontrar um vetor x que satisfaça exatamente a equação $Ax = b$, pelo que estamos interessamos

no problema de encontrar o vetor x (ordem m) que minimiza o erro quadrático dado por:

$$E := \sum_{i=1}^{n} [z_i - b_i]^2 \tag{6.1}$$

onde $z = Ax$ e z_i é linha i do vetor z, dado por:

$$z_i = (Ax)_i = \sum_{j=1}^{m} a_{ij} x_j, \quad i = 1, \cdots, n \tag{6.2}$$

onde a_{ij} é o elemento de A na linha i e coluna j. Substituindo (6.2) em (6.1)

$$E := \sum_{i=1}^{n} \left[\sum_{j=1}^{m} a_{ij} x_j - b_i \right]^2 \tag{6.3}$$

Esta é uma função diferenciável nos coeficientes x_j e portanto todo ponto de mínimo acontece quando $\nabla E = 0$, ou seja, quando

$$\frac{\partial}{\partial x_l} E = 0, \forall 1 \leq l \leq m$$

O que implica a seguinte condição

$$0 = \frac{\partial}{\partial x_l} E = \sum_{i=1}^{n} 2 \left[\sum_{j=1}^{m} a_{ij} x_j - b_i \right] a_{il}, \quad l = 1, \cdots, m$$

Equivalente a

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_{il} x_j a_{ij} = \sum_{i=1}^{n} a_{il} b_i, \quad l = 1, \cdots, m$$

que pode ser reescrito na forma vetorial como:

$$\begin{bmatrix} \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i1} x_j a_{ij} \\ \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i2} x_j a_{ij} \\ \vdots \\ \sum_{i=1}^{n} \sum_{j=1}^{m} a_{im} x_j a_{ij} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} a_{i1} b_i \\ \sum_{i=1}^{n} a_{i2} b_i \\ \vdots \\ \sum_{i=1}^{n} a_{im} b_i \end{bmatrix} \tag{6.4}$$

Observamos agora que a expressão (6.4) é equivalente ao seguinte problema matricial:

$$\boxed{A^T A x = A^T b} \tag{6.5}$$

Teorema 5. *A matriz $M = A^T A$ é quadrada de ordem m e é invertível sempre que o posto da matriz A é igual a número de colunas m.*

Demonstração. Para provar que M é invertível precisamos mostrar que $Mv = 0$ implica $v = 0$:

$$Mv = 0 \Longrightarrow A^T Av = 0$$

tomando o produto interno da expressão $0 = A^T Av$ com v, temos:

$$0 = \left\langle A^T Av, v \right\rangle = \langle Av, Av \rangle = \|Av\|^2$$

Então se $Mv = 0$ $Av = 0$, como o posto de A é igual ao número de colunas, $v = 0$. □

Outra propriedade importante é que M é simétrica, ou seja, $M = M^T$. Isso é facilmente provado pelo seguinte argumento:

$$M^T = (A^T A)^T = (A)^T (A^T)^T = A^T A = M$$

6.6.2 Ajuste linear de curvas pelo método dos mínimos quadrados

Seja $f_1(x), f_2(x), \ldots, f_m(x)$ um conjunto de m funções e $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ um conjunto de n pontos. Procuram-se os coeficientes $a_1, a_2, \ldots, a_m$ tais que a função dada por

$$f(x) = a_1 f_1(x) + a_2 f_2(x) + \ldots + a_m f_m(x)$$

minimiza o erro dado por

$$E_q = \sum_{i=1}^{n} [f(x_i) - y_i]^2$$

como $f(x) = \sum_{j=1}^{m} a_j f_j(x)$, temos

$$E_q = \sum_{i=1}^{n} \left[\sum_{j=1}^{m} a_j f_j(x_i) - y_i \right]^2$$

Este problema é equivalente a resolver pelo métodos dos mínimos quadrados o seguinte sistema linear:

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_m(x_2) \\ f_1(x_3) & f_2(x_3) & \cdots & f_m(x_3) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_m(x_n) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

Exemplo 51. *Encontrar a reta que melhor aproxima o seguinte conjunto de dados:*

x_i	y_i
0,01	*1,99*
1,02	*4,55*
2,04	*7,2*
2,95	*9,51*
3,55	*10,82*

Desejamos então encontrar os valores de a e b tais que a função $f(x) = ax+b$ *melhor se ajusta aos pontos da tabela. Afim de usar o critério dos mínimos quadrados, escrevemos o problema na forma matricial dada por:*

$$\begin{bmatrix} 0,01 & 1 \\ 1,02 & 1 \\ 2,04 & 1 \\ 2,95 & 1 \\ 3,55 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1,99 \\ 4,55 \\ 7,2 \\ 9,51 \\ 10,82 \end{bmatrix}$$

Multiplicamos agora ambos os lados pela transposta $\begin{bmatrix} 0,01 & 1,02 & 2,04 & 2,95 & 3,55 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$*:*

$$\begin{bmatrix} 0,01 & 1,02 & 2,04 & 2,95 & 3,55 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0,01 & 1 \\ 1,02 & 1 \\ 2,04 & 1 \\ 2,95 & 1 \\ 3,55 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0,01 & 1,02 & 2,04 & 2,95 & 3,55 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix}$$

$$\begin{bmatrix} 26,5071 & 9,57 \\ 9,57 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 85,8144 \\ 34,07 \end{bmatrix}$$

A solução desse sistema é $a = 2,5157653$ *e* $b = 1,9988251$

A tabela abaixo mostra os valores dados e os valores ajustados:

x_i	y_i	$ax_i + b$
0,01	1,99	2,0239828
1,02	4,55	4,5649057
2,04	7,2	7,1309863
2,95	9,51	9,4203327
3,55	10,82	10,929792

Problema 18. *Encontrar a parábola* $y = ax^2 + bx + c$ *que melhor aproxima o seguinte conjunto de dados:*

x_i	y_i
$0,01$	$1,99$
$1,02$	$4,55$
$2,04$	$7,2$
$2,95$	$9,51$
$3,55$	$10,82$

e complete a tabela

x_i	y_i	$ax_i^2 + bx_i + c$	$ax_i^2 + bx_i + c - y_i$
0,01	*1,99*		
1,02	*4,55*		
2,04	*7,2*		
2,95	*9,51*		
3,55	*10,82*		

Resposta $y = -0,0407898x^2 + 2,6613293x + 1,9364598$

x_i	y_i	$ax_i^2 + bx_i + c$	$ax_i^2 + bx_i + c - y_i$
0,01	1,99	1,963069	-0,0269310
1,02	4,55	4,6085779	0,0585779
2,04	7,2	7,1958206	-0,0041794
2,95	9,51	9,4324077	-0,0775923
3,55	10,82	10,870125	0,0501249

Problema 19. *Dado o seguinte conjunto de dados*

x_i	y_i
0,0	*31*
0,1	*35*
0,2	*37*
0,3	*33*
0,4	*28*
0,5	*20*
0,6	*16*
0,7	*15*
0,8	*18*
0,9	*23*
1,0	*31*

- *Encontre a função do tipo* $f(x) = a + b\sin(2\pi x) + c\cos(2\pi x)$ *que melhor aproxima os valores dados.*

- *Encontre a função do tipo $f(x) = a+bx+cx^2+dx^3$ que melhor aproxima os valores dados.*

Resp: $a = 25,638625$, $b = 9,8591874$, $c = 4,9751219$ e $a = 31,475524$, $b = 65,691531$, $c = -272,84382$, $d = 208,23621$.

6.7 Problemas não lineares que podem ser aproximados por problemas lineares

Eventualmente, problemas de ajuste de curvas podem recair num sistema não linear. Por exemplo, se desejamos ajustar a função $y = Ae^{bx}$ ao conjunto de pontos (x_0, y_0), (x_1, y_1) e (x_2, y_2), temos que minimizar o funcional

$$E_q = (Ae^{x_0 b} - y_0)^2 + (Ae^{x_1 b} - y_1)^2 + (Ae^{x_2 b} - y_2)^2$$

ou seja, resolver o sistema

$$\begin{aligned} \frac{\partial E_q}{\partial A} &= 2(Ae^{x_0 b} - y_0)e^{x_0 b} + 2(Ae^{x_1 b} - y_1)e^{x_1 b} + 2(Ae^{x_2 b} - y_2)e^{x_2 b} = 0 \\ \frac{\partial E_q}{\partial b} &= 2Ax_0(Ae^{x_0 b} - y_0)e^{x_0 b} + 2Ax_1(Ae^{x_1 b} - y_1)e^{x_1 b} + 2x_2 A(Ae^{x_2 b} - y_2)e^{x_2 b} = 0 \end{aligned}$$

que é não linear em A e b. Esse sistema pode ser resolvido pelo método de Newton-Raphson, o que pode se tornar custoso, ou mesmo inviável quando não dispomos de uma boa aproximação da solução para inicializar o método.

Felizmente, algumas famílias de curvas admitem uma transformação que nos leva a um problema linear. No caso da curva $y = Ae^{bx}$, observe que $\ln y = \ln A + bx$. Assim, em vez de ajustar a curva original $y = Ae^{bx}$ a tabela de pontos, ajustamos a curva submetida a transformação logarítmica

$$z = \ln A + bx := B + bx.$$

Usamos os três pontos $(x_0, \ln y_0) := (x_0, \tilde{y}_0)$, $(x_1, \ln y_1) := (x_1, \tilde{y}_1)$ e $(x_2, \ln y_2) := (x_2, \tilde{y}_2)$ e resolvemos o sistema linear

$$A^T A \begin{bmatrix} B \\ b \end{bmatrix} = A^T \begin{bmatrix} \tilde{y}_0 \\ \tilde{y}_1 \\ \tilde{y}_2 \end{bmatrix},$$

onde

$$A = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

Exemplo 52. *Encontre uma curva da forma* $y = Ae^x$ *que melhor ajusta os pontos* $(1, 2)$, $(2, 3)$ *e* $(3, 5)$.

Temos

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

e a solução do sistema leva em $B = 0,217442$ e $b = 0,458145$. Portanto, $A = e^{0,217442} = 1,24289$.

Observação 21. *Os coeficientes obtidos a partir dessa linearização são aproximados, ou seja, são diferentes daqueles obtidos quando aplicamos mínimos quadrados não linear. Observe que estamos minimizando* $\sum_i [\ln y_i - \ln(f(x_i))]^2$ *em vez de* $\sum_i [y_i - f(x_i)]^2$. *No exemplo resolvido, a solução do sistema não linear original seria* $A = 1,19789$ *e* $B = 0,474348$

Observação 22. *Mesmo quando se deseja resolver o sistema não linear, a solução do problema linearizado pode ser usada para construir condições iniciais.*

A próxima tabela apresenta algumas curvas e transformações que linearizam o problema de ajuste.

curva	transformação	problema linearizado
$y = ae^{bx}$	$Y = \ln y$	$Y = \ln a + bx$
$y = ax^b$	$Y = \ln y$	$Y = \ln a + b \ln x$
$y = ax^b e^{cx}$	$Y = \ln y$	$Y = \ln a + b \ln x + cx$
$y = ae^{(b+cx)^2}$	$Y = \ln y$	$Y = \ln a + b^2 + bcx + c^2x^2$
$y = \frac{a}{b+x}$	$Y = \frac{1}{y}$	$Y = \frac{b}{a} + \frac{1}{a}x$
$y = A\cos(\omega x + \phi)$ ω conhecido	–	$y = a\cos(\omega x) - b\sin(\omega x)$, $a = A\cos(\phi)$, $b = A\sin(\phi)$

Exemplo 53. *Encontre a função* f *da forma* $y = f(x) = A\cos(2\pi x + \phi)$ *que*

ajusta a tabela de pontos

x_i	y_i
0,0	*9,12*
0,1	*1,42*
0,2	*- 7,76*
0,3	*- 11,13*
0,4	*- 11,6*
0,5	*- 6,44*
0,6	*1,41*
0,7	*11,01*
0,8	*14,73*
0,9	*13,22*
1,0	*9,93*

Usando o fato que $y = A\cos(2\pi x + \phi) = a\cos(2\pi x) - b\sin(2\pi x)$, onde $a = A\cos(\phi)$ e $b = A\sin(\phi)$, $z = [\ a \quad b\ ]^T$ é solução do problema

$$B^T B z = B^T y,$$

onde

$$B = \begin{bmatrix} \cos(2\pi x_0) & -\sin(2\pi x_0) \\ \cos(2\pi x_1) & -\sin(2\pi x_1) \\ \vdots & \\ \cos(2\pi x_{10}) & -\sin(2\pi x_{10}) \end{bmatrix} = \begin{bmatrix} 1. & 0. \\ 0,8090170 & -0,5877853 \\ 0,3090170 & -0,9510565 \\ -0,3090170 & -0,9510565 \\ -0,8090170 & -0,5877853 \\ -1,0000000 & 0,0000000 \\ -0,8090170 & 0,5877853 \\ -0,3090170 & 0,9510565 \\ 0,3090170 & 0,9510565 \\ 0,8090170 & 0,5877853 \\ 1,0000000 & 0,0000000 \end{bmatrix}.$$

Assim, $a = 7,9614704$ e $b = 11,405721$ e obtemos o seguinte sistema:

$$\begin{cases} A\cos(\phi) = 7,9614704 \\ A\sin(\phi) = 11,405721 \end{cases}.$$

Observe que

$$A^2 = 7,9614704^2 + 11,405721^2$$

e, escolhendo $A > 0$, $A = 13,909546$ e

$$\sin(\phi) = \frac{11,405721}{13,909546} = 0,8199923$$

Assim, como $\cos\phi$ também é positivo, ϕ é um ângulo do primeiro quadrante:

$$\phi = 0,9613976$$

Portanto $f(x) = 13,909546\cos(2\pi x + 0,9613976)$. Observe que nesse exemplo a solução do problema linear é a mesma do problema não linear.

Problema 20. *Encontre a função f da forma $y = f(x) = \frac{a}{b+x}$ que ajusta a tabela de pontos*

x_i	y_i
0,0	*101*
0,2	*85*
0,4	*75*
0,6	*66*
0,8	*60*
1,0	*55*

usando uma das transformações tabeladas.

Usando o fato que $Y = \frac{1}{y} = \frac{b}{a} + \frac{1}{a}x$, $z = [\ \frac{b}{a} \quad \frac{1}{a}\ ]^T$ é solução do problema

$$A^T A z = A^T Y,$$

onde

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \end{bmatrix} = \begin{bmatrix} 1 & 0,0 \\ 1 & 0,2 \\ 1 & 0,4 \\ 1 & 0,6 \\ 1 & 0,8 \\ 1 & 1,0 \end{bmatrix}$$

e

$$Y = \begin{bmatrix} 1/y_1 \\ 1/y_2 \\ 1/y_3 \\ 1/y_4 \\ 1/y_5 \\ 1/y_6 \end{bmatrix} = \begin{bmatrix} 0,0099010 \\ 0,0117647 \\ 0,0133333 \\ 0,0151515 \\ 0,0166667 \\ 0,0181818 \end{bmatrix}$$

Assim, $\frac{1}{a} = 0,0082755$ e $\frac{b}{a} = 0,0100288$ e, então, $a = 120,83924$ e $b = 1,2118696$, ou seja, $f(x) = \frac{120,83924}{1,2118696+x}$.

6.8 Interpolação linear segmentada

Considere o conjunto $(x_i, y_i)_{j=1}^{n}$ de n pontos. Assumiremos que $x_{i+1} > x_i$, ou seja, as abscissas são distintas e estão em ordem crescente. A função linear que interpola os pontos x_i e x_{i+1} no intervalo i é dada por

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

O resultado da interpolação linear segmentada é a seguinte função contínua definida por partes no intervalo $[x_1, x_n]$:

$$f(x) = P_i(x), \quad x \in [x_i, x_{i+1}]$$

Exemplo 54. *Construa uma função linear por partes que interpola os pontos* $(0,0)$, $(1,4)$, $(2,3)$, $(3,0)$, $(4,2)$, $(5,0)$.

A função procurada pode ser construída da seguinte forma:

$$f(x) = \begin{cases} 0\frac{x-1}{0-1} + 1\frac{x-0}{1-0} & , 0 \leq x < 1 \\ 4\frac{x-2}{1-2} + 3\frac{x-1}{2-1} & , 1 \leq x < 2 \\ 3\frac{x-3}{2-3} + 0\frac{x-2}{3-2} & , 2 \leq x \leq 3 \end{cases}$$

Simplificando, obtemos:

$$f(x) = \begin{cases} x & , 0 \leq x < 1 \\ -x + 5 & , 1 \leq x < 2 \\ -3x + 9 & , 2 \leq x \leq 3 \end{cases}$$

A Figura 6.2 é um esboço da função $f(x)$ obtida. Ela foi gerada no Scilab usando os comandos:

```
//pontos fornecidos
xi = [0;1;2;3;4;5]
yi = [0;4;3;0;2;0]
//numero de pontos
n = 6
//funcao interpoladora
function [y] = f(x)
  for i=1:n-2
    if ((x>=xi(i)) & (x<xi(i+1))) then
      y = yi(i)*(x-xi(i+1))/(xi(i) - xi(i+1)) ...
          + yi(i+1)*(x-xi(i))/(xi(i+1) - xi(i));
    end
```

```
  end

  if ((x>=xi(n-1)) & (x<=xi(n))) then
    y = yi(n-1)*(x-xi(n))/(xi(n-1) - xi(n)) ...
        + yi(n)*(x-xi(n-1))/(xi(n) - xi(n-1));
  end
endfunction
//graficando
xx = linspace(xi(1),xi(n),500)';
clear yy
for i=1:max(size(xx))
  yy(i) = f(xx(i))
end
plot(xi,yi,'r.',xx,yy,'b-')
```

Veja a Figura 6.2.

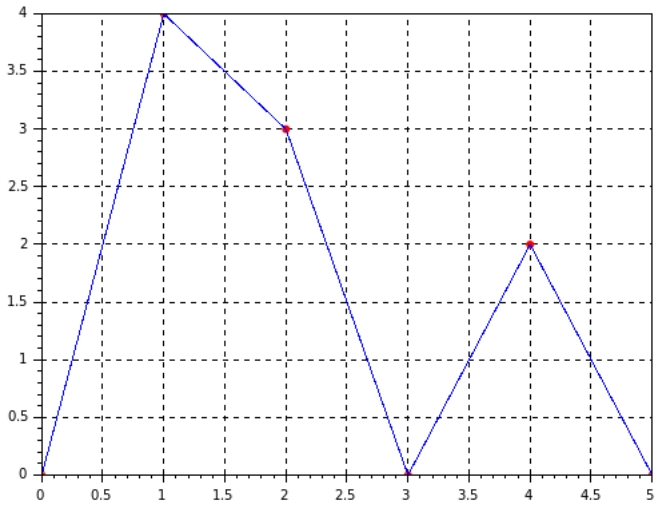


Figura 6.2: Interpolação linear segmentada.

6.9 Interpolação cúbica segmentada - spline

Dado um conjunto de n pontos $(x_j, y_j)_{j=1}^{n}$ tais que $x_{j+1} > x_j$, ou seja, as abscissas são distintas e estão em ordem crescente; um spline cúbico que

interpola estes pontos é uma função $s(x)$ com as seguintes propriedades:

i Em cada segmento $[x_j, x_{j+1}]$, $j = 1, 2, \ldots n-1$ $s(x)$ é um polinômio cúbico.

ii para cada ponto, $s(x_j) = y_j$, i.e., o spline interpola os pontos dados.

iii $s(x) \in C^2$, i.e., é função duas vezes continuamente diferenciável.

Da primeira hipótese, escrevemos

$$s(x) = s_j(x), x \in [x_j, x_{j+1}], \quad j = 1, \ldots, n-1$$

com

$$s_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

O problema agora consiste em obter os 4 coeficientes de cada um desses $n-1$ polinômios cúbicos.

Veremos que a simples definição de spline produz $4n-6$ equações linearmente independentes:

$$\begin{array}{rcll} s_j(x_j) & = & y_j, & j = 1, \ldots, n-1 \\ s_j(x_{j+1}) & = & y_{j+1}, & j = 1, \ldots, n-1 \\ s'_j(x_{j+1}) & = & s'_{j+1}(x_{j+1}), & j = 1, \ldots, n-2 \\ s''_j(x_{j+1}) & = & s''_{j+1}(x_{j+1}), & j = 1, \ldots, n-2 \end{array}$$

Como

$$s'_j(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2 \tag{6.6}$$

e

$$s''_j(x) = 2c_j + 6d_j(x - x_j), \tag{6.7}$$

temos, para $j = 1, \ldots, n-1$, as seguintes equações

$$\begin{array}{rcl} a_j & = & y_j, \\ a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3 & = & y_{j+1}, \\ b_j + 2c_j(x_{j+1} - x_j) + 3d_j(x_{j+1} - x_j)^2 & = & b_{j+1}, \\ c_j + 3d_j(x_{j+1} - x_j) & = & c_{j+1}, \end{array}$$

Por simplicidade, definimos

$$h_j = x_{j+1} - x_j$$

e temos

$$\begin{array}{rcl} a_j & = & y_j, \\ a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 & = & y_{j+1}, \\ b_j + 2c_j h_j + 3d_j h_j^2 & = & b_{j+1}, \\ c_j + 3d_j h_j & = & c_{j+1}, \end{array}$$

que podem ser escrita da seguinte maneira

$$a_j = y_j, \tag{6.8}$$

$$d_j = \frac{c_{j+1} - c_j}{3h_j}, \tag{6.9}$$

$$\begin{aligned} b_j &= \frac{y_{j+1} - y_j - c_j h_j^2 - \frac{c_{j+1}-c_j}{3h_j} h_j^3}{h_j}, \\ &= \frac{3y_{j+1} - 3y_j - 3c_j h_j^2 - c_{j+1} h_j^2 + c_j h_j^2}{3h_j} \\ &= \frac{3y_{j+1} - 3y_j - 2c_j h_j^2 - c_{j+1} h_j^2}{3h_j} \end{aligned} \tag{6.10}$$

Trocando o índice j por $j-1$ na terceira equação (6.8), $j = 2, \ldots, n-1$

$$b_{j-1} + 2c_{j-1}h_{j-1} + 3d_{j-1}h_{j-1}^2 = b_j \tag{6.11}$$

e, portanto,

$$\begin{aligned} \frac{3y_j - 3y_{j-1} - 2c_{j-1}h_{j-1}^2 - c_j h_{j-1}^2}{3h_{j-1}} + 2c_{j-1}h_{j-1} + c_j h_{j-1} - c_{j-1}h_{j-1} \\ = \frac{3y_{j+1} - 3y_j - 2c_j h_j^2 - c_{j+1} h_j^2}{3h_j}. \end{aligned} \tag{6.12}$$

Fazendo as simplificações, obtemos:

$$c_{j-1}h_{j-1} + c_j(2h_j + 2h_{j-1}) + c_{j+1}h_j = 3\frac{y_{j+1} - y_j}{h_j} - 3\frac{y_j - y_{j-1}}{h_{j-1}}. \tag{6.13}$$

É costumeiro acrescentar a incógnita c_n ao sistema. A incógnita c_n não está relacionada a nenhum dos polinômios interpoladores. Ela é uma construção artificial que facilita o cálculo dos coeficientes do spline. Portanto, a equação acima pode ser resolvida para $j = 2, \ldots, n-1$.

Para determinar unicamente os n coeficientes c_n precisamos acrescentar duas equações linearmente independentes às $n-2$ equações dadas por (6.13). Essas duas equações adicionais definem o tipo de spline usado.

6.9.1 Spline natural

Uma forma de definir as duas equações adicionais para completar o sistema (6.13) é impor condições de fronteira livres (ou naturais), ou seja,

$$S''(x_1) = S''(x_n) = 0. \tag{6.14}$$

Substituindo na equação (6.7)

$$s_1''(x_1) = 2c_1 + 6d_1(x_1 - x_1) = 0 \Longrightarrow c_1 = 0.$$

e

$$s_{n-1}''(x_n) = 2c_{n-1} + 6d_{n-1}(x_n - x_{n-1}) = 0.$$

Usando o fato que

$$c_{n-1} + 3d_{n-1}h_{n-1} = c_n$$

temos que

$$c_n = -3d_{n-1}(x_n - x_{n-1}) + 3d_{n-1}h_{n-1} = 0.$$

Essa duas equações para c_1 e c_n juntamente com as equações (6.13) formam um sistema de n equações $Ac = z$, onde

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ h_1 & 2h_2 + 2h_1 & h_2 & 0 & \cdots & 0 & 0 \\ 0 & h_2 & 2h_3 + 2h_2 & h_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-2} & 2h_{n-2} + 2h_{n-1} & h_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{6.15}$$

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \quad \text{e} \quad z = \begin{bmatrix} 0 \\ 3\frac{y_3 - y_2}{h_2} - 3\frac{y_2 - y_1}{h_1} \\ 3\frac{y_4 - y_3}{h_3} - 3\frac{y_3 - y_2}{h_2} \\ \vdots \\ 3\frac{y_{n-1} - y_{n-2}}{h_{n-2}} - 3\frac{y_{n-2} - y_{n-3}}{h_{n-3}} \\ 0 \end{bmatrix} \tag{6.16}$$

Observe que a matriz A é diagonal dominante estrita e, portanto, o sistema $Ac = z$ possui solução única. Calculado c, os valores dos a_n, b_n e d_n são obtidos diretamente pelas expressões (6.8), (6.10) e (6.9), respectivamente.

Exemplo 55. *Construa um spline cúbico natural que passe pelos pontos* $(2,\ 4,5)$, $(5,-1,9)$, $(9,\ 0,5)$ *e* $(12,-0,5)$.

O spline desejado é uma função definida por partes da forma:

$$f(x) = \begin{cases} a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3 & , 2 \leq x < 5 \\ a_2 + b_2(x-5) + c_2(x-5)^2 + d_2(x-5)^3 & , 5 \leq x < 9 \\ a_3 + b_3(x-9) + c_3(x-9)^2 + d_3(x-9)^3 & , 9 \leq x \leq 12 \end{cases} . \tag{6.17}$$

Os coeficientes c_1, c_2 e c_3 resolvem o sistema $Ac = z$, onde

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 2\cdot 3 + 2\cdot 4 & 4 & 0 \\ 0 & 4 & 2\cdot 4 + 2\cdot 3 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 14 & 4 & 0 \\ 0 & 4 & 14 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} \qquad \text{e} \qquad z = \begin{bmatrix} 0 \\ 3\frac{0,5-(-1,9)}{4} - 3\frac{(-1,9)-4,5}{3} \\ 3\frac{-0,5-0,5}{3} - 3\frac{0,5-(-1,9)}{4} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 8,2 \\ -2,8 \\ 0 \end{bmatrix}$$

Observe que c_4 é um coeficiente artificial para o problema. A solução é $c_1 = 0$, $c_2 = 0,7$, $c_3 = -0,4$ e $c_4 = 0$. Calculamos os demais coeficientes usando as expressões (6.8), (6.10) e (6.9):

$$\begin{aligned} a_1 &= y_1 = 4,5 \\ a_2 &= y_2 = -1,9 \\ a_3 &= y_3 = 0,5 \end{aligned}$$

$$\begin{aligned} d_1 &= \frac{c_2 - c_1}{3h_1} = \frac{0,7-0}{3\cdot 3} = 0,0777778 \\ d_2 &= \frac{c_3 - c_2}{3h_2} = \frac{-0,4-0,7}{3\cdot 4} = -0,0916667 \\ d_3 &= \frac{c_4 - c_3}{3h_3} = \frac{0+0,4}{3\cdot 3} = 0,0444444 \end{aligned}$$

$$\begin{aligned} b_1 &= \frac{y_2 - y_1}{h_1} - \frac{h_1}{3}(2c_1 + c_2) = \frac{-1,9-4,5}{3} - \frac{3}{3}(2\cdot 0 - 0,7) = -2,8333333 \\ b_2 &= \frac{y_3 - y_2}{h_2} - \frac{h_2}{3}(2c_2 + c_3) = \frac{0,5-(-1,9)}{4} - \frac{4}{3}(2\cdot 0,7 + 0,4) = -0,7333333 \\ b_3 &= \frac{y_4 - y_3}{h_3} - \frac{h_3}{3}(2c_3 + c_4) = \frac{-0,5-0,5}{3} - \frac{3}{3}(2\cdot(-0,4) + 0) = 0,4666667 \end{aligned}$$

Portanto,

$$f(x) = \begin{cases} 4,5 - 2,8333333(x-2) + 0,0777778(x-2)^3 & ,2 \le x < 5 \\ -1,9 - 0,7333333(x-5) + 0,7(x-5)^2 - 0,0916667(x-5)^3 & ,5 \le x < 9 \\ 0,5 + 0,4666667(x-9) - 0,4(x-9)^2 + 0,0444444(x-9)^3 & ,9 \le x \le 12 \end{cases}.$$

No Scilab, podemos utilizar:

```
X = [2 5 9 12]'
Y = [4.5 -1.9 0.5 -0.5]'
h = X(2:4)-X(1:3)
A = [1 0 0 0;h(1) 2*h(1)+2*h(2) h(2) 0; ...
   0 h(2) 2*h(2)+2*h(3) h(3);0 0 0 1 ]
z = [0, 3*(Y(3)-Y(2))/h(2)-3*(Y(2)-Y(1))/h(1), ...
   3*(Y(4)-Y(3))/h(3)-3*(Y(3)-Y(2))/h(2), 0]'
c = A\z
for i=1:3
   a(i) = Y(i)
   d(i) = (c(i+1)-c(i))/(3*h(i))
   b(i) = (Y(i+1)-Y(i))/h(i)-h(i)/3*(2*c(i)+c(i+1))
end

for i=1:3
    P(i) = poly([a(i) b(i) c(i) d(i)],'x','coeff')
    z = [X(i):.01:X(i+1)]
    plot(z,horner(P(i),z-X(i)))
end
```

6.9.2 Spline com condições de contorno fixadas

Alternativamente, para completar o sistema (6.13), podemos impor condições de contorno fixadas, ou seja,

$$\begin{aligned} S'(x_1) &= f'(x_1) \\ S'(x_n) &= f'(x_n). \end{aligned}$$

Substituindo na equação (6.6)

$$s_1'(x_1) = b_1 + 2c_1(x_1 - x_1) + 3d_j(x_1 - x_1)^2 = f'(x_1) \Longrightarrow b_1 = f'(x_1) \quad (6.18)$$

e

$$\begin{aligned} s_{n-1}'(x_n) &= b_{n-1} + 2c_{n-1}(x_n - x_{n-1}) + 3d_j(x_n - x_{n-1})^2 \\ &= b_{n-1} + 2c_{n-1}h_{n-1} + 3d_{n-1}h_{n-1}^2 = f'(x_n) \end{aligned} \quad (6.19)$$

Usando as equações (6.9) e (6.10) para $j = 1$ e $j = n - 1$, temos:

$$2c_1h_1 + c_2h_1 = 3\frac{y_2 - y_1}{h_1} - 3f'(x_1) \quad (6.20)$$

e

$$c_{n-1}h_{n-1} + c_nh_{n-1} = 3f'(x_n) - 3\frac{y_n - y_{n-1}}{h_{n-1}} \quad (6.21)$$

Essas duas equações juntamente com as equações (6.13) formam um sistema de n equações $Ac = z$, onde

$$A = \begin{bmatrix} 2h_1 & h_1 & 0 & 0 & \cdots & 0 & 0 \\ h_1 & 2h_2+2h_1 & h_2 & 0 & \cdots & 0 & 0 \\ 0 & h_2 & 2h_3+2h_2 & h_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-2} & 2h_{n-2}+2h_{n-1} & h_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix} \tag{6.22}$$

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \quad \text{e} \quad z = \begin{bmatrix} 3\frac{y_2-y_1}{h_1} - 3f'(x_1) \\ 3\frac{y_3-y_2}{h_2} - 3\frac{y_2-y_1}{h_1} \\ 3\frac{y_4-y_3}{h_3} - 3\frac{y_3-y_2}{h_2} \\ \vdots \\ 3\frac{y_{n-1}-y_{n-2}}{h_{n-2}} - 3\frac{y_{n-2}-y_{n-3}}{h_{n-3}} \\ 3f'(x_n) - 3\frac{y_n-y_{n-1}}{h_{n-1}} \end{bmatrix} \tag{6.23}$$

Observe que a matriz A é diagonal dominante estrita e, portanto, o sistema $Ac = z$ possui solução única. Calculado c, os valores dos a_n, b_n e d_n são obtidos diretamente pelas expressões (6.8), (6.10) e (6.9), respectivamente.

Exemplo 56. *Construa um spline cúbico com fronteira fixada que interpola a função $y = \sin(x)$ nos pontos $x = 0$, $x = \frac{\pi}{2}$, $x = \pi$, $x = \frac{3\pi}{2}$ e $x = 2\pi$.*

O spline desejado passa pelos pontos $(0, 0)$, $(\pi/2, 1)$, $(\pi, 0)$, $(3\pi/2, -1)$ e $(2\pi, 0)$ e tem a forma:

$$f(x) = \begin{cases} a_1 + b_1x + c_1x^2 + d_1x^3 & , 0 \le x < \frac{\pi}{2} \\ a_2 + b_2(x-\frac{\pi}{2}) + c_2(x-\frac{\pi}{2})^2 + d_2(x-\frac{\pi}{2})^3 & , \frac{\pi}{2} \le x < \pi \\ a_3 + b_3(x-\pi) + c_3(x-\pi)^2 + d_3(x-\pi)^3 & , \pi \le x < \frac{3\pi}{2} \\ a_4 + b_4(x-\frac{3\pi}{2}) + c_4(x-\frac{3\pi}{2})^2 + d_4(x-\frac{3\pi}{2})^3 & , \frac{3\pi}{2} \le x \le 2\pi \end{cases} . \tag{6.24}$$

Observe que ele satisfaz as condição de contorno $f'(0) = cos(0) = 1$ e $f'(2\pi) = cos(2\pi) = 1$.

Os coeficientes c_1, c_2, c_3 e c_4 resolvem o sistema $Ac = z$, onde:

$$A = \begin{bmatrix} \pi & \pi/2 & 0 & 0 & 0 \\ \pi/2 & 2\pi & \pi/2 & 0 & 0 \\ 0 & \pi/2 & 2\pi & \pi/2 & 0 \\ 0 & 0 & \pi/2 & 2\pi & \pi/2 \\ 0 & 0 & 0 & \pi/2 & \pi \end{bmatrix} \tag{6.25}$$

$$c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} \quad \text{e} \quad z = \begin{bmatrix} 3\frac{1-0}{\pi/2} - 3 \cdot 1 \\ 3\frac{0-1}{\pi/2} - 3\frac{1-0}{\pi/2} \\ 3\frac{-1-0}{\pi/2} - 3\frac{0-1}{\pi/2} \\ 3\frac{0-(-1)}{\pi/2} - 3\frac{(-1)-0}{\pi/2} \\ 3 \cdot 1 - 3\frac{0-(-1)}{\pi/2} \end{bmatrix} = \begin{bmatrix} 6/\pi - 3 \\ -12/\pi \\ 0 \\ 12/\pi \\ 3 - 6/\pi \end{bmatrix} \tag{6.26}$$

Aqui c_5 é um coeficiente artificial para o problema. A solução é $c_1 = -0,0491874$, $c_2 = -0,5956302$, $c_3 = 0$, $c_4 = 0,5956302$ e $c_5 = 0,0491874$. Calculamos os demais coeficientes usando as expressões (6.8), (6.10) e (6.9):

$$\begin{aligned} a_1 &= y_1 = 0 \\ a_2 &= y_2 = 1 \\ a_3 &= y_3 = 0 \\ a_4 &= y_3 = -1 \end{aligned}$$

$$\begin{aligned} d_1 &= \frac{c_2 - c_1}{3h_1} = \frac{-0,5956302 - (-0,0491874)}{3 \cdot \pi/2} = -0,1159588 \\ d_2 &= \frac{c_3 - c_2}{3h_2} = \frac{0 - (-0,5956302)}{3 \cdot \pi/2} = 0,1263967 \\ d_3 &= \frac{c_4 - c_3}{3h_3} = \frac{0,5956302 - 0}{3 \cdot \pi/2} = 0,1263967 \\ d_4 &= \frac{c_5 - c_4}{3h_4} = \frac{0,0491874 - 0,5956302}{3 \cdot \pi/2} = -0,1159588 \end{aligned}$$

$$b_1 = \frac{y_2 - y_1}{h_1} - \frac{h_1}{3}(2c_1 + c_2) = \frac{1-0}{\pi/2} - \frac{\pi/2}{3}(2 \cdot (-0,0491874) - 0,5956302) = 1$$

$$b_2 = \frac{y_3 - y_2}{h_2} - \frac{h_2}{3}(2c_2 + c_3) = \frac{0-1}{\pi/2} - \frac{\pi/2}{3}(2 \cdot (-0,5956302) + 0) = -0,0128772$$

$$b_3 = \frac{y_4 - y_3}{h_3} - \frac{h_3}{3}(2c_3 + c_4) = \frac{-1-0}{\pi/2} - \frac{\pi/2}{3}(2 \cdot 0 + 0,5956302) = -0,9484910$$

$$b_4 = \frac{y_5 - y_4}{h_4} - \frac{h_4}{3}(2c_4 + c_5) = \frac{0-(-1)}{\pi/2} - \frac{\pi/2}{3}(2 \cdot 0,5956302 + 0,0491874) = -0,0128772$$

Portanto,

$$f(x) = \begin{cases} x - 0,0491874x^2 - 0,1159588x^3 & , 0 \leq x < \frac{\pi}{2} \\ 1 + -0,0128772(x - \frac{\pi}{2}) - 0,5956302(x - \frac{\pi}{2})^2 + 0,1263967(x - \frac{\pi}{2})^3 & , \frac{\pi}{2} \leq x < \pi \\ -0,9484910(x - \pi) + 0,1263967(x - \pi)^3 & , \pi \leq x < \frac{3\pi}{2} \\ -1 - 0,0128772(x - \frac{3\pi}{2}) + 0,5956302(x - \frac{3\pi}{2})^2 - 0,1159588(x - \frac{3\pi}{2})^3 & , \frac{3\pi}{2} \leq x \leq 2 \end{cases}$$

No Scilab, podemos resolver este problema fazendo:

Resumo sobre Splines

Dado um conjunto de pontos (x_i, y_i), $i = 1, 2, \ldots, n$, um spline cúbico é a seguinte função definida por partes:

$$s(x) = \begin{cases} s_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3 & , x_1 \leq x < x_2 \\ s_2(x) = a_2 + b_2(x - x_2) + c_2(x - x_2)^2 + d_2(x - x_2)^3 & , x_2 \leq x < x_3 \\ \quad \vdots & \quad \vdots \\ s_{n-1} = a_{n-1} + b_{n-1}(x - x_{n-1}) + c_{n-1}(x - x_{n-1})^2 + d_{n-1}(x - x_{n-1})^3 & , x_{n-1} \leq x \leq x_n \end{cases}$$

Definindo-se $h_j = x_{j+1} - x_j$, os coeficientes c_j, $j = 1, 2, \ldots, n$, são solução do sistema linear $Ac = B$, onde:

Spline Natural $s_1''(x_1) = 0$ e $s_{n-1}''(x_n) = 0$	Spline Fixado $s_1'(x_1) = f'(x_1)$ e $s_{n-1}'(x_n) = f'(x_n)$
$a_{i,j} = \begin{cases} 1 & , j = i = 1 \\ h_{i-1} & , j = i - 1, i < n \\ 2(h_i + h_{i-1}) & , j = i, 1 < i < n \\ h_i & , j = i + 1, i > 1 \\ 1 & , j = i = n \\ 0 & , \text{caso contrário.} \end{cases}$	$a_{i,j} = \begin{cases} 2h_1 & , j = i = 1 \\ h_{i-1} & , j = i - 1 \\ 2(h_i + h_{i-1}) & , j = i, 1 < i < n \\ h_i & , j = i + 1 \\ 2h_{n-1} & , j = i = n \\ 0 & , \text{caso contrário.} \end{cases}$
$b_i = \begin{cases} 0 & , i = 1 \\ 3\frac{y_{i+1} - y_i}{h_i} - 3\frac{y_i - y_{i-1}}{h_{i-1}} & , 1 < i < n \\ 0 & , i = n \end{cases}$	$b_i = \begin{cases} 3\frac{y_2 - y_1}{h_1} - 3f'(x_1) & , i = 1 \\ 3\frac{y_{i+1} - y_i}{h_i} - 3\frac{y_i - y_{i-1}}{h_{i-1}} & , 1 < i < n \\ 3f'(x_n) - 3\frac{y_n - y_{n-1}}{h_{n-1}} & , i = n \end{cases}$

os coeficientes a_j, b_j e d_j, $j = 1, 2, \ldots, n - 1$, são calculados conforme segue:

$$a_j = y_j$$

$$b_j = \frac{3y_{j+1} - 3y_j - 2c_j h_j^2 - c_{j+1} h_j^2}{3h_j}$$

$$d_j = \frac{c_{j+1} - c_j}{3h_j}$$

Referências Bibliográficas

[1] Cecill and free sofware. `http://www.cecill.info`. Acessado em 30 de julho de 2015.

[2] M. Baudin. Introduction to scilab. `http://forge.scilab.org/index.php/p/docintrotoscilab/`. Acessado em 30 de julho de 2015.

[3] R.L. Burden and J.D. Faires. *Análise Numérica*. Cengage Learning, 8 edition, 2013.

[4] J. P. Demailly. *Analyse Numérique et Équations Differentielles*. EDP Sciences, Grenoble, nouvelle Édition edition, 2006.

[5] Walter Gautschi and Gabriele Inglese. Lower bounds for the condition number of vandermonde matrices. *Numerische Mathematik*, 52(3):241–250, 1987/1988.

[6] R. Rannacher. Einführung in die numerische mathematik (numerik 0). `http://numerik.uni-hd.de/~lehre/notes/num0/numerik0.pdf`. Acessado em 10.08.2014.