Notas da disciplina Cálculo Numérico

Leonardo F. Guidi

29 de novembro de 2017

Instituto de Matemática
Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500
Porto Alegre - RS

Sumário

1 Representação de números em máquinas

1.1 Sistema de numeração

Um sistema de numeração é formado por uma coleção de símbolos e regras para representar conjuntos de números de maneira consistente. Um sistema de numeração que desempenhe satisfatoriamente o seu propósito deve possuir as seguintes propriedades:

1. Capacidade de representar um conjunto de números distintos de maneira padronizada.
2. Deve refletir as estruturas algébricas e aritméticas dos números

O sistema mais utilizado é o *sistema de numeração posicional de base* 10 ou *sistema de numeração base*-10 – o sistema decimal. Em um sistema de numeração posicional a posição relativa dos símbolos guarda informação sobre o número que se quer representar. Mais especificamente, posições adjacentes estão relacionadas entre si por uma constante multiplicativa denominada *base* do sistema de numeração. Assim, cada símbolo em uma determinada posição contribui aditivamente com uma quantidade dada pela multiplicação do valor numérico do símbolo pela base elevada ao expoente dado pelo posição. Dessa forma, o *numeral*, formado por uma sequência de símbolos, representa o número calculado a partir da soma da contribuição conjunta de cada símbolo e de sua posição.

Definição 1.1.1 (**Sistemas de numeração base-*b***). *Sistemas de numeração base-b. Dado um natural*[1] $b > 1$, *a coleção de símbolos "−", "," e os algarismos*[2] $\{0, 1, \ldots, b-1\}$, *o numeral* [3]

$$d_n d_{n-1} \cdots d_1 d_0 , d_{-1} \cdots {}_b \tag{1.1.1}$$

representa o número real positivo

$$\sum_{j \leq n} d_j b^j . \tag{1.1.2}$$

Aplicando o sinal "−" à frente do numeral representamos os reais negativos.

Observação 1.1.2. *Quando não há possibilidade de equívoco, dispensamos o subscrito* 10 *na representação decimal. Dessa forma, por exemplo, o numeral* $1{,}23_{10}$, *ou simplesmente* $1{,}23$, *representa o número* $1 \times 10^0 + 2 \times 10^{-1} + 3 \times 10^{-2}$. *Os algarismos* 0 *antes da vírgula também não*

[1] É possível construir sistemas de numeração cujas bases são inteiros negativos, números irracionais e até com números complexos. Em particular, é possível construir sistemas de numeração binários com base $1 + i$ ou $1 - i$ (proposta por Walter F. Penney em 1965). Esse tipo de sistema é capaz de representar os números complexos e pode ser implementada em hardware, o leitor pode consultar o artigo:

- Jamil, T "The complex binary number system". *IEEE Potentials* **20**: 39–41 (2002).

[2] Quando a base é maior do que 10 utilizamos as letras maiúsculas $A, B, C, \ldots$ para representar os algarismos – essa é a notação utilizada na base hexadecimal (base-16).

[3] Por se tratar de números reais, a sequência sempre é limitada à esquerda e portanto $n \in \mathbb{Z}$.

são representados se não houver mais algum outro algarismo à esquerda. O mesmo ocorre para os algarismos zero à direita após a vírgula com exceção do uso nas ciências exatas quando faz-se necessário registrar a precisão de alguma medida. Assim, evitamos notações da forma $00\ldots02{,}3$ *para indicar o racional* $\dfrac{23}{10}$*; o mesmo para notações da forma* $2{,}30\ldots00$*, com exceção dos casos em que se quer indicar não o racional* $\dfrac{23}{10}$ *mas uma medida com determinada precisão. Quando a representação de um número possuir dízimas periódicas utiliza-se como notação uma barra sobre a sequência que se repete. Por exemplo,* $\frac{1}{3}$ *é representado como* $0{,}\overline{3}$ *e* $\dfrac{67}{495}$ *é representado como* $0{,}1\overline{35}$.

Os algarismos à esquerda da vírgula formam a *parte inteira* do numeral, os demais formam a sua *parte fracionária*. Os sistemas de numeração base-b. O sistemas de numeração base-b são capazes de representar unicamente os números inteiros, no entanto, o mesmo não é verdade para números racionais. Por exemplo, o racional $\dfrac{11}{10}$ possui as representações $1{,}1$ e $1{,}0\overline{9}$. Essa ambiguidade na representação posicional se deve ao fato de que para uma base $b > 1$ e um $k \in \mathbb{Z}$

$$\sum_{j=-\infty}^{k} (b-1)\, b^j = b^{k+1}.$$

Uma forma de evitar a ambiguidade é exigir na definição do sistema de numeração posicional base-b que no numeral (1.1.1) deve sempre haver uma quantidade infinita de dígitos não nulos d_j que satisfaçam a desigualdade $d_j < b - 1$. Neste caso, representações da forma $1{,}0\overline{9}$ não são admitidas e em seu lugar utiliza-se a representação $1{,}1$[4].

1.1.1 Mudança de base

A partir das expressões (1.1.1) e (1.1.2) é possível construir o procedimento de mudança de bases. Seja, então, X um número representado na base b como $d_n d_{n-1} \cdots d_1 d_0, d_{-1} \ldots {}_b$. Encontrar a sua representação em uma outra base g significa encontrar a sequência de algarismos $\tilde{d}_m \tilde{d}_{m-1} \cdots \tilde{d}_1 \tilde{d}_0, \tilde{d}_{-1} \ldots {}_g$ tal que

$$X = \sum_{j \leq m} \tilde{d}_j g^j.$$

A seguinte abordagem consiste em separar X nas suas as partes inteira e fracionária. Seja X^i a parte inteira de X,

$$X^i = \sum_{j=0}^{m} \tilde{d}_j g^j. \tag{1.1.3}$$

De maneira semelhante, seja X^f a parte fracionária de X,

$$X^f = \sum_{j \leq -1} \tilde{d}_j g^j. \tag{1.1.4}$$

Sem perda de generalidade, a parte inteira será tratada inicialmente. A divisão de X^i por g dá

[4] Agradeço aos professores William R. P. Conti e Domingos H. U. Marchetti por essa observação.

origem a dois termos:

$$\frac{X^i}{g} = \frac{\tilde{d}_0}{g} + \sum_{j=1}^{m} \tilde{d}_j g^{j-1},$$

o primeiro termo no lado direito da expressão anterior é uma fração, o segundo é um número inteiro. Entendendo a operação anterior como uma divisão ineira, pode-se dizer então que a divisão de X^i por g (realizada na base b) possui $\sum_{j=1}^{m} \tilde{d}_j g^{j-1}$ (que estará representado na base b em que a operação foi realizada) como quociente e $\tilde{d}_0$ (também na base b) como resto. Portanto, essa operação de divisão inteira nos fornece o primeiro dígito da nova representação, $\tilde{d}_0$. Em seguida, a divisão do segundo membro da igualdade anterior (ou seja, o quociente da divisão anterior) fornece $\tilde{d}_1$:

$$\frac{\sum_{j=1}^{m} \tilde{d}_j g^{j-1}}{g} = \frac{\tilde{d}_1}{g} + \sum_{j=2}^{m} \tilde{d}_j g^{j-2}.$$

E assim sucessivamente até ser obtido o dígito $\tilde{d}_m$ quando o quociente da última divisão for nulo. Convém lembrar que ao iniciar o processo, apenas a representação de X^i na base b é conhecida, portanto **as operações de divisão devem ser feitas na base** b. Em resumo, o procedimento incia-se com a representação de X^ina base b, em seguida obtém-se a representação de g na base b. Uma vez determinada essa representação, X^i é divido por g; o resto da divisão é o primeiro algarismo da representação de X^i em base g – o algarismo em si está representado na base b – e o quociente deverá ser dividido novamente no passo seguinte.

Ao contrário do que ocorre no procedimento para a mudança de base na parte inteira, a mudança de base para a parte fracionária envolve operações de multiplicação. A partir da expressão (1.1.4) é possível observar que

$$g\, X^f = \tilde{d}_{-1} + \sum_{j=-2}^{\infty} \tilde{d}_j g^{j+1},$$

o que fornece o primeiro dígito da parte fracionária – note que no lado direito da expressão anterior o primeiro termo é um inteiro e o segundo é um fracionário. Repetindo a operação com a parte fracionária de $g\, X^f$:

$$g \sum_{j\leq -2} \tilde{d}_j g^{j+1} = \tilde{d}_{-2} + \sum_{j\leq -3} \tilde{d}_j g^{j+2}.$$

E assim por diante, os termos da parte fracionária de X na base g são obtidos. Aqui vale a mesma observação do parágrafo anterior, as operações de multiplicação devem ser realizadas na base b e os dígitos obtidos estão representados na base b.

Exemplo 1: Vamos representar o numeral $53{,}2\overline{05}_6$ no sistema base-8. Nesse caso $X^i = 53_6$ e $X^f = 0{,}2\overline{05}_6$. Para encontrar os dígitos da parte inteira na base 8 devemos realizar sucessivas operações de divisão por 8, como X^i está representado na base 6, devemos realizar todas as

operações nessa base[5], ou seja vamos dividir por $12_6 (= 8_{10})$:

$$\frac{53_6}{12_6} = 4_6 + \frac{1_6}{12_6},$$

o resto da divisão, 1_6 é o primeiro dígito (na base 6), na base 8 temos o mesmo dígito, ou seja, 1_8. Em seguida, vamos dividir o quociente 4_6:

$$\frac{4_6}{12_6} = 0_6 + \frac{4_6}{12_6},$$

aqui o resto da divisão é $4_6 = 4_8$ e o procedimento termina pois o quociente da divisão é nulo. Portanto a parte inteira é representada pelo numeral 41_8.
Encontramos a representação da parte fracionária através de operações de multiplicação por 8 na base 6, ou seja, 12_6:

$$0{,}2\overline{05}_6 \times 12_6 = 2{,}\overline{50}_6 = 2_6 + 0{,}\overline{50}_6 \ ,$$

a parte inteira da multiplicação[6], 2_6, é o primeiro dígito após a vírgula em base 8, 2_8. Em seguida multiplicamos a parte fracionária $0{,}\overline{50}_6$ por 8, ou melhor, 12_6:

$$0{,}\overline{50}_6 \times 12_6 = 10{,}\overline{50}_6 = 10_6 + 0{,}\overline{50}_6 \ ,$$

a parte inteira que resulta da multiplicação é $10_6 = 6_8$ e a parte fracionária novamente assume a mesma forma, $0{,}\overline{50}_6$. Isto nos permite concluir que sempre obteremos o mesmo dígito a partir deste ponto. Portanto, $0{,}2\overline{05}_6$ representa o mesmo número que $0{,}2\overline{6}_8$. Combinando a parte inteira e fracionária temos finalmente que $53{,}2\overline{05}_6 = 41{,}2\overline{6}_8$.

[5] Para facilitar as operações é conveniente utilizar uma tabela de multiplicações em base 6 (tabuada em base 6):

	1_6	2_6	3_6	4_6	5_6
1_6	1_6	2_6	3_6	4_6	5_6
2_6	2_6	4_6	10_6	12_6	14_6
3_6	3_6	10_6	13_6	20_6	23_6
4_6	4_6	12_6	20_6	24_6	32_6
5_6	5_6	14_6	23_6	32_6	41_6

[6] Aqui cabe uma observação sobre o cálculo de produtos envolvendo dízimas periódicas: o produto $0{,}2\overline{05}_6 \times 12_6$ pode ser decomposto como a soma

$$\begin{aligned} 0{,}2\overline{05}_6 \times 10_6 + 0{,}2\overline{05}_6 \times 2_6 &= 2{,}\overline{05}_6 + \left(0{,}2\overline{05}_6 \times 2_6\right) \\ &= 2{,}\overline{05}_6 + \left(\left(0{,}2_6 + 0{,}0\overline{05}_6\right) \times 2_6\right) \\ &= 2{,}\overline{05}_6 + \left(0{,}4_6 + 0{,}0\overline{05}_6 \times 2_6\right). \end{aligned}$$

Por sua vez, o produto dentro do parênteses, $0{,}0\overline{05}_6 \times 2_6$, corresponde ao somatório $2_6 \times \sum_{k=1}^{\infty} 5_6 \times 6^{-(2k+1)}$. Como $2_6 \times 5_6 = 14_6$, temos que $0{,}0\overline{05}_6 \times 2_6 = \sum_{k=1}^{\infty} 14_6 \times 6^{-(2k+1)}$, ou seja, $0{,}0\overline{05}_6 \times 2_6 = 0{,}0\overline{14}_6$ e portanto $0{,}2\overline{05}_6 \times 2_6 = 0{,}4\overline{14}_6$. Assim o resultado do produto $0{,}2\overline{05}_6 \times 12_6$ é igual à soma $2{,}\overline{05}_6 + 0{,}4\overline{14}_6$.

Somas cujos termos possuem dízimas periódicas são calculadas da forma usual, no entanto é necessário uma maior atenção no transporte de algarismo para as posições seguintes. Neste caso, $2{,}\overline{05}_6 + 0{,}4\overline{14}_6 = 2{,}0\overline{50}_6 + 0{,}4\overline{14}_6$ e a soma dos algarismos 5_6 e 1_6 nas posições pares da dízima acarreta o transporte de 1_6 para as posições ímpares. Dessa forma, temos que $2{,}\overline{05}_6 + 0{,}4\overline{14}_6 = 2{,}\overline{50}_6$. A seguir, esse resultado é apresentado em detalhe com o uso de

1.1.2 Bits e bytes (e nibbles também...)

O termo "bit" é uma contração de "binary digit" – dígito binário – e como tal, pode assumir dois valores distintos. Registrado pela primeira vez no artigo "Instrumental Analysis" do engenheiro americano Vannevar Bush, publicado[7] em 1936. O termo "byte", criado em 1956 pelo cientista da computação Werner Buchholz durante a fase de desenvolvimento do primeiro supercomputador transistorizado IBM 7030, também conhecido como IBM Stretch, correspondia originalmente um conjunto de tamanho variável formado por um a oito bits. Devido à popularidade da arquitetura IBM System/360 (anos 1960) e da família de microprocessadores de 8 bits lançados nos anos 1970, a quantia universalmente aceita é de oito bits para cada byte. O termo "nibble" não possui uma origem clara. Atualmente é utilizado para representar um grupo de quatro bits, ou seja, meio byte.

Na representação posicional base-2, o bit assume os valores 0 ou 1. Nas máquinas, um bit é a menor unidade de informação; seus dois possíveis valores podem ser interpretados como "verdadeiro" ou "falso" o que possibilita a construção de operações lógicas em máquina. São as operações lógicas que permitem a realização das operações aritméticas com os registros das máquinas.

Um numeral em representação binária com quatro dígitos

$$(d_3 d_2 d_1 d_0)_2$$

correspondem à mesma informação que pode ser armazenada por um nibble. Em se tratando de um inteiro sem sinal, os valores no intervalo $[0, 15]$. Por sua vez, esses valores correspondem ao algarismos do sistema base-16 (base hexadecimal), ou seja cada nibble corresponde a um algarismo hexadecimal.

---

somatórios para representar as dízimas:

$$\begin{aligned}
0{,}2\overline{05}_6 \times 12_6 &= 2{,}\overline{05}_6 + 0{,}4\overline{14}_6 \\
&= 2_6 + \left(\sum_{k=1}^{\infty} 5_6 \times 6^{-2k}\right) + 0{,}4_6 + \left(\sum_{k=1}^{\infty} 1_6 \times 6^{-2k}\right) + \left(\sum_{k=1}^{\infty} 4_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}4_6 + \left(\sum_{k=1}^{\infty} (5_6 + 1_6) \times 6^{-2k}\right) + \left(\sum_{k=1}^{\infty} 4_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}4_6 + \left(\sum_{k=1}^{\infty} 10_6 \times 6^{-2k}\right) + \left(\sum_{k=1}^{\infty} 4_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}4_6 + \left(\sum_{k=1}^{\infty} 1_6 \times 6^{-2k+1}\right) + \left(\sum_{k=1}^{\infty} 4_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}4_6 + \left(0{,}1_6 + \sum_{k=1}^{\infty} 1_6 \times 6^{-(2k+1)}\right) + \left(\sum_{k=1}^{\infty} 4_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}5_6 + \left(\sum_{k=1}^{\infty} (1_6 + 4_6) \times 6^{-(2k+1)}\right) \\
&= 2{,}5_6 + \left(\sum_{k=1}^{\infty} 5_6 \times 6^{-(2k+1)}\right) \\
&= 2{,}\overline{50}_6.
\end{aligned}$$

[7] Bush, V. "Instrumental Analysis", Bull. Amer. Math. Soc. **42** (1936), 649-669.

binário	hexadecimal
0000_2	0_{16}
0001_2	1_{16}
0010_2	2_{16}
0011_2	3_{16}
0100_2	4_{16}
0101_2	5_{16}
0110_2	6_{16}
0111_2	7_{16}
1000_2	8_{16}
1001_2	9_{16}
1010_2	A_{16}
1011_2	B_{16}
1100_2	C_{16}
1101_2	D_{16}
1110_2	E_{16}
1111_2	F_{16}

A partir dessa correspondência, uma maneira conveniente e natural de representar um sequência formada por múltiplos inteiros de 8 bits é através de sequências equivalentes de bytes representados por pares de algarismos hexadecimais. O desenvolvimento a seguir ilustra essa possibilidade.

A sequência de 8 bits $d_7d_6d_5d_4d_3d_2d_1d_0$ pode ser entendida como a representação binária do número inteiro

$$d_7 \times 2^7 + d_6 \times 2^6 + d_5 \times 2^5 + d_4 \times 2^4 + d_3 \times 2^3 + d_2 \times 2^2 + d_1 \times 2 + d_0.$$

Ao reagrupar os termos da soma na forma

$$\left(d_7 \times 2^3 + d_6 \times 2^2 + d_5 \times 2 + d_4\right) \times 2^4 + \left(d_3 \times 2^3 + d_2 \times 2^2 + d_1 \times 2 + d_0\right)$$

é possível verificar que os algarimos d_0 a d_3 e os algarimos d_4 a d_7 compõe, cada um, um nibble:

$$\left(d_7d_6d_5d_4\right)_2 \times 16 + \left(d_3d_2d_1d_0\right)_2 .$$

Sejam $h_1 = \left(d_7d_6d_5d_4\right)_2$ e $h_0 = \left(d_3d_2d_1d_0\right)_2$ dois algarismos hexadecimais. A soma anterior corresponde a um número cuja representação hexadecimal é da forma $\left(h_1h_0\right)_{16}$. A sequência de algarismos hexadecimais h_1h_0 é entendida como um byte. Por exemplo, a sequência de 16 bits 0001 1111 1000 1011 é convenientemente representada pelos 2 bytes 1F 8B.

1.2 Aritmética de máquina

No ocidente, a utilização de um sistema posicional base-10 com algarismos indo-arábicos é corrente desde pelo menos o século XIV, no entanto, em outras culturas é comum encontrarmos

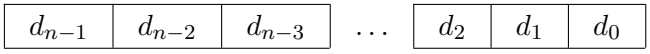
sistemas – ou pelo menos o seu reflexo na linguagem – base-5, base-8, base-12 e mesmo a utilização matemática de sistemas posicionais base-20 (civilização maia) e base-60 utilizada pelos sumérios e babilônios com reflexos até hoje na notação para medir ângulos – em graus minutos e segundos – e nas unidades de tempo minuto e segundo[8].

Tipicamente, um número inteiro é armazenado em um processador como uma sequência de dígitos binários de comprimento fixo denominada registro. Os processadores dispõe de um ou mais circuitos integrados denominados ALUs (plural de unidade lógica e aritmética[9]) cujo papel é realizar esse tipo de operações nos registros.

Os computadores digitais atuais, em quase sua totalidade, utilizam ALUs que representam internamente os números em base-2 (base binária) e/ou base-10 e realizam operações aritméticas nessas bases[10]. No início da computação eletrônica chegaram a ser construídas máquinas que representavam números em base ternária (base-3)[11]. Apesar de sua maior eficência e menor custo de fabricação, o desenvolvimento foi interrompido devido à crescente produção (e consequente barateamento devido à economia de escala) e desenvolvimento de componentes para a construção de processadores binários.

1.2.1 Representação de números inteiros

Se todo o registro for utilizado para representar um inteiro não negativo a representação é única: um registro de n bits da forma



[8] O leitor mas interessado deve ler o fascinante texto de Donald Knuth:

- Knuth, D. E. "The Art of Computer Programming, vol2. Seminumerical Algorithms", 3ª edição. Addison-Wesley, 1997.

em particular, o início do capítulo 4 sobre aritmética.

[9] Em lígua inglesa, *arithmetic and logic unit* que dá origem ao acrônimo. As ALUs foram conceitualmente propostas em 1945 por John von Neumann com parte do computador EDVAC (*Electronic Discrete Variable Automatic Computer*), um dos primeiros computadores eletrônicos binários.

[10] A imensa maioria dos processadores atuais possuem regsitros e realizam operações em base binária. A exceção são os processadores utilizados em calculadoras científicas e os seguintes processadores fabricados pela IBM: POWER6, unidades de processamento do System z9 e System z10. Esses processadores dispõe de ALUs que permitem o registro e operação em base-10.

[11] O único exemplo é o computador SETUN desenvolvido em 1958 na Universidade Estatal de Moscou Lomnossov por Sergei Sobolev e Nikolay Brusentsov. Essas máquinas foram criadas até 1965 e um novo modelo foi desenvolvido em 1970, o SETUN-70. Detalhes podem ser obtidos em:

- Klimenko, Stanislav V.: Computer science in Russia: A personal view. IEEE Annals of the history of computing, v 21, n 3, 1999.
- Žogolev, Y. A.: The order code and an interpretative system for the Setun computer. USSR Comp. Math. And Math. Physics (3), 1962, Oxford, Pergamon Press, p 563-578.
- G. Trogemann, A. Y. Nitussov, W. Ernst (Hg.), Computing in Russia: The History of Computer Devices and Information Technology revealed. Vieweg Verlag, July 2001.
- Hunger, Francis: SETUN. An Inquiry into the Soviet Ternary Computer. Institut für Buchkunst Leipzig, 2008, ISBN 3-932865-48-0.

representa o número $(d_{n-1}d_{n-2}d_{n-3}\ldots d_2d_1d_0)_2$. Assim, é possível representar os números inteiros entre 0 , representado por $000\ldots000_2$ até o inteiro representado por $111\ldots111_2$:

$$111\ldots111_2 = 2^{n-1} + 2^{n-2} + 2^{n-3} + \cdots + 2^2 + 2^1 + 2^0 = 2^n - 1.$$

Naturalmente, também é necessário representar também os inteiros com sinal. No entanto, existem diferentes maneiras de representá-los através de um registro binário. A seguir, três das maneiras mais comuns: representação com bit de sinal, representação complemento de um e a representação complemento de dois.

Representação com bit de sinal

Essa forma de representação foi utilizada nas ALUs dos primeiro computadores digitais binários produzidos comercialmente[12]. Consiste na utilização de um dos bits para o sinal. Geralmente, o bit mais significativo no registro (o primeiro à esquerda) é utilizado para esse fim.

De acordo com essa representação, um registro de máquina formado por uma sequência de 32 bits :

s	d_{30}	d_{29}	$\ldots$	d_2	d_1	d_0

é utilizado para representar o número binário $((-1)^s d_{30}\, d_{29}\, \ldots\, d_2\, d_1\, d_0)_2$. Ou seja, a sequência binária de 32 dígitos (ou bits) $s\, d_{30}\, d_{29}\, \ldots\, d_2\, d_1\, d_0$ é capaz de representar todos os inteiros entre $-2^{31}+1$ e $2^{31}-1$ (-2147483647 e 2147483647) na forma

$$(-1)^s \left(d_0 2^0 + d_1 2^1 + \ldots + d_{30} 2^{30}\right).$$

Nesse caso o maior inteiro representável é dado numeral $I_{max} = 011\ldots11_2$. Portanto

$$I_{max} = (-1)^0 \left(1 + 2^1 + 2^2 + \ldots + 2^{30}\right) = 2^{31} - 1 = 2.147.483.647.$$

Levando em conta os números negativos e que o zero possui duas representações possíveis (como 0 e -0), essa disposição de dados no registro permite representar os $2^{32}-1$ inteiros entre $-\left(2^{31}-1\right)$ e $2^{31}-1$.

Essa técnica de registro com n bits permite armazenar em uma máquina todos os 2^n-1 números inteiros entre $-\left(2^{n-1}-1\right)$ e $2^{n-1}-1$.

Se em uma operação de soma ou subtração o resultado for um número que não pode ser armazenado nos registros ocorre um erro conhecido como *overflow*. Nesse caso a máquina deve ser capaz de reconhecer o evento e enviar uma mensagem de erro – se não o fizesse, poderia retornar um número truncado que não corresponde ao resultado correto da operação programada.

[12] Um dos primeiros modelos de computador/maninframe fabricados pela IBM a partir de 1956, o IBM305 RAMAC – um computador no qual ainda eram empregadas válvulas – utilizava essa forma de representação de inteiros com sinal. Um dos primeiros modelos transistorizados, o IBM1401 lançado em 1959, também utilizava essa representação.

Representação complemento de um

Consiste na utilização de um registro de n bits da seguinte forma: o primeiro bit representa um termo aditivo $-\left(2^{n-1}-1\right)$ e o restante dos bits representam um inteiro não negativo. Assim um registro n bits da forma

d_{n-1}	d_{n-2}	d_{n-3}	$\ldots$	d_2	d_1	d_0

representa o inteiro $(d_{n-2}d_{n-3}\ldots d_2d_1d_0)_2 - d_{n-1}\left(2^{n-1}-1\right)$.

O nome da representação advém do fato de que nessa representação, os dígitos do inverso aditivo de um número são obtidos através do seu complemento reduzido de base[13] no sistema binário. Essa operação equivale também a tomar os complementos reduzidos de base dos dígitos separadamente.

Essa representação contém os inteiros entre $-\left(2^{n-1}-1\right)$ e $2^{n-1}-1$. O zero possui também duas representações: $111\ldots11_2$ e $000\ldots00_2$.

Exemplos:

O registro de 8 bits

0	1	0	0	0	1	1	0

representa o inteiro $1000110_2 - 0\left(2^7-1\right) = 70$.

O registro de 8 bits

1	0	1	1	1	0	0	1

representa o inteiro $0111001_2 - 1\left(2^7-1\right) = 57 - 127 = -70$. Comparando os dois registros, podemos notar que o segundo é obtido a partir do primeiro trocando-se os dígitos 1 por 0 e vice-versa[14]. Essa forma de representação também foi utilizada em ALUs de computadores mais antigos[15].

Representação complemento de dois

A representação complemento de dois é a representação mais utilizada nas ALUs dos processadores atuais por garantir uma maior simplicidade no desenho de circuitos para operações lógicas e aritméticas. Consiste na utilização de um registro de n bits da seguinte forma: o primeiro bit representa um termo aditivo -2^{n-1} e o restante dos bits representam um inteiro não negativo. Assim um registro n bits da forma

d_{n-1}	d_{n-2}	d_{n-3}	$\ldots$	d_2	d_1	d_0

representa o inteiro $(d_{n-2}d_{n-3}\ldots d_2d_1d_0)_2 - d_{n-1}2^{n-1}$.

[13] O "complemento reduzido de base" de um inteiro y com n dígitos no sistema base-b é definido como $(b^n-1)-y$. Assim, no sistema binário, o complemento reduzido de base de 0 é 1 e vice-versa.

[14] Ou seja, é a operação lógica "não" aplicada a cada bit.

[15] Entre eles o PDP-1 lançado em 1960 pela Digital Equipament Corporation, um computador muito utilizado nos departamentos de ciências exatas de várias universidades (como o MIT, onde em 1961 Steve Russell, Martin Graetz and Wayne Wiitanen apresentaram o primeiro *game* programado em um computador digital, o "Spacewar!"); e o modelo 160A da Control Data Corporation lançado em 1960 a partir do projeto desenvolvido por Seymor Cray (anos mais tarde ele fundaria a Cray Corporation, famosa por seus supercomputadores).

Nessa representação, o registro do inverso aditivo de um número é obtido a partir do seu complemento de base[16].

Essa representação contém os inteiros entre -2^{n-1} e $2^{n-1}-1$ e ao contrário das anteriores, o zero é unicamente representado por $000\ldots00_2$.

Exemplos:

O registro de 8 bits

0	1	0	0	0	1	1	0

representa o inteiro $1000110_2 - 0\left(2^7\right) = 70$.

O registro de 8 bits

1	0	1	1	1	0	1	0

representa o inteiro $0111010_2 - 1\left(2^7\right) = 58 - 128 = -70$. Comparando os dois registros, podemos notar que o registro do inverso aditivo de um número é obtido trocando-se os dígitos 1 por 0 e vice-versa e adicionando 1 ao resultado.

1.2.2 Representação de números com parte fracionária – ponto–fixo

É possível estender a técnica utilizada na representação dos inteiros para representar com precisão finita, números que possuam parte fracionária. Dados dois inteiros positivos p e q, interpretamos um registro de $n = p + q + 1$ bits como a divisão por 2^q do inteiro de n bits na representação complemento de dois. Vamos simbolizar tal registro como $R(p, q)$.

Dessa forma, um registro de 32 bits $R(15, 16)$:

d_{31}	d_{30}	d_{29}	$\ldots$	d_2	d_1	d_0

representa o número

$$\left((d_{30}d_{29}\ldots d_2d_1d_0)_2 - d_{31}2^{31}\right)2^{-16} \longrightarrow \underbrace{\left((d_{30}d_{29}\ldots d_{17}d_{16})_2 - d_{31}2^{15}\right)}_{\text{parte inteira}} + \underbrace{(0{,}d_{15}d_{14}\ldots d_1d_0)_2}_{\text{parte fracionária}}$$

Esse tipo de representação de números é conhecido com *representação de ponto-fixo*. De forma geral, utilizando $p + 1$ bits para representar a parte inteira e q bits para a parte fracionária é possível representar os números fracionários no intervalo $[-2^p, 2^p - 2^{-q}]$ em intervalos igualmente espaçados de 2^{-q}. Em módulo, o menor número representável também possui esse valor. Se em alguma operação o resultado for um número menor em módulo que 2^{-q} diz-se que ocorreu um erro de *underflow*, em particular a região compreendida pelo intervalo $(-2^{-q}, 2^{-q})$ é denominada *região de underflow*. Da mesma maneira, se o resultado de alguma operação for maior que $2^p - 2^{-q}$ ou menor do que -2^p, diz-se que ocorreu um erro de *overflow* e a região $(-\infty, -2^p) \cup (2^p - 2^{-q}, +\infty)$ é denominada *região de overflow*.

A representação em ponto fixo possui a vantagem de oferecer uma representação para números com parte fracionária que podem ser trabalhados dentro de ALUs, ou seja, é possível realizar operações aritméticas com números não inteiros através de circuitos integrados de fabricação simples. Por esse motivo ela é utilizada em sistemas onde a simplicidade de fabricação e operação são

[16] O "complemento de base" de um inteiro y com n dígitos no sistema base-b é definido como $b^n - y$.

fundamentais.

Entretanto, os registros de ponto fixo possuem a desvantagem de representar números distintos com precisão diferente. Por exemplo os números $9999{,}\overline{1234}$ e $0{,}0012\overline{13}$ são representados em base 10 com 4 dígitos para a parte inteira e 4 dígitos para a parte fracionária da seguinte forma: 9999,1234 e 0,0012. Enquanto que no primeiro caso o número é representado com oito dígitos, o segundo dispõe de apenas dois dígitos para representá-lo. Essa assimetria na representação em ponto fixo, caracterizada pela maior precisão com que os números de maior valor absoluto são registrados, motiva a introdução da representação de ponto flutuante.

1.2.3 Representação de ponto flutuante

Definição 1.2.1 (ponto flutuante). *A representação x de um número real é denominada ponto flutuante normalizado na base $b \in \mathbb{N}$, $b \geq 2$, se forem satisfeitas as propriedades*

1. $x = m\, b^{\varepsilon}$, *onde*
2. $m = \pm d_1{,}d_2 \ldots d_n \qquad n \in \mathbb{N}$,
3. $1 \leq d_1 \leq b-1 \quad e \quad 0 \leq d_i \leq b-1 \quad$ *para* $i = 2, 3, \ldots n$,
4. $e_1 \leq \varepsilon \leq e_2, \quad$ *onde* $\varepsilon,\ e_1,\ e_2 \in \mathbb{Z}$.

m é denominada significando[17]*, ε expoente e n o número de dígitos de precisão.*

> **Exemplo 2:** O número $9999{,}\overline{1234}$ em representação de ponto flutuante em base 10 com 8 dígitos de precisão é $9{,}9991234 \cdot 10^3$. Utilizando essa mesma prescrição, o número $0{,}0012\overline{13}$ é representado como $1{,}2113131 \cdot 10^{-3}$.

Em uma representação de ponto flutuante normalizado, o primeiro algarismo após a vírgula é necessariamente maior ou igual a 1. Portanto o número zero está fora dos casos cobertos pela definição de ponto flutuante. Usualmente o incluímos em um conjunto denominado *sistema de ponto flutuante* onde possui a representação $0{,}00 \ldots 0\, b^{\varepsilon}$. O sistema de ponto flutuante $F(b, n, e_1, e_2)$ é definido como o conjunto de números que inclui o zero e os pontos flutuantes em base b com n dígitos de precisão e expoente que pode variar entre e_1 e e_2 inclusive.

Propriedades do conjunto $F(b, n, e_1, e_2)$

Ao contrário do que ocorre com os números representados por um esquema de ponto fixo, os elementos sucessivos do conjunto $F(b, n, e_1, e_2)$ não são igualmente espaçados. Para exemplificar essa propriedade vamos considerar o sistema dado pelo conjunto $F(10, 2, -10, 10)$. De acordo com a definição, o elemento positivo mais próximo de zero é o numeral $1{,}0 \cdot 10^{-10}$, o numeral seguinte é $1{,}1 \cdot 10^{-10}$ e assim por diante até o numeral $9{,}9 \cdot 10^{-10}$. O espaçamento entre eles[18] é de $0{,}1 \cdot 10^{-10}$. Após o numeral $9{,}9 \cdot 10^{-10}$, vêm os numerais $1{,}0 \cdot 10^{-9}$, $1{,}1 \cdot 10^{-9}$, $\ldots$, $9{,}9 \cdot 10^{-9}$. Agora

[17] Também conhecido como coeficiente, ou ainda, mantissa.

[18] Note que a diferença entre esses primeiros numerais é menor do que o menor numeral representável pelo sistema, ou seja, se for realizada uma operação de subtração em ponto flutuante entre quaisquer dois elementos consecutivos o resultado será nulo.

o espaçamento já é $0{,}1 \cdot 10^{-9}$ e assim por diante até os maiores numerais $1{,}0 \cdot 10^{10}, \ldots, 9{,}9 \cdot 10^{10}$ cujo espaçamento é $0{,}1 \cdot 10^{10}$. Portanto os elementos estão mais densamente acumulados em torno do zero.

A cardinalidade de um sistema de ponto flutuante $F(b, n, e_1, e_2)$ é dada por $|F(b, n, e_1, e_2)|$:

$$|F(b, n, e_1, e_2)| = 1 + 2(b-1)b^{n-1}(e_2 - e_1 + 1). \tag{1.2.1}$$

A cardinalidade do conjunto é calculada a partir de todas as combinações possíveis para a representação de um numeral como elemento de $F(b, n, e_1, e_2)$, somadas ao elemento zero que não pode ser representado segundo a definição de ponto flutuante usual. O primeiro termo do lado direito de (1.2.1) deve-se ao zero. O fator 2 deve-se ao sinal. O fator $(b-1)$ deve-se aos possíveis valores que o dígito d_1 pode assumir. O fator b^{n-1} deve-se à combinação dos b possíveis valores que os dígitos $d_2, \ldots, d_n$ podem assumir. E finalmente, o fator $(e_2 - e_1 + 1)$ deve-se aos possíveis valores que o expoente pode assumir.

Arredondamento

A operação de arredondamento consiste em encontrar um representação $\tilde{x}$ para um número x com uma determinada precisão. Essa operação é usualmente realizada em máquinas para representar internamente os números e o resultado de operações aritméticas realizadas sobre eles. Não há uma única forma de realizar o arredondamento e, de acordo com a aplicação, existem regras mais convenientes. Vamos discutir as cinco mais comuns e previstas pelo padrão da IEEE.

Qualquer que seja a operação de arredondamento, o resultado depende exclusivamente dos dígitos que compõe o número, a sua magnitude não desempenha papel algum. Assim, por simplicidade, o tratamento exposto considerará números da forma

$$x = \pm(d_0{,}d_{-1}\ldots)_b.$$

Uma operação de arredondamento do número x com k dígitos resulta em um número $\tilde{x}$ com k dígitos de precisão

$$\tilde{x} = \pm(\tilde{d}_0{,}\tilde{d}_{-1}\ldots\tilde{d}_{-k+1})_b$$

onde os valores dos dígitos $\tilde{d}_0$, $\tilde{d}_{-1}$, $\ldots$, $\tilde{d}_{-k+1}$ são determinados a partir dos valores de d_{-k+1}, d_{-k}, d_{-k-1}, $\ldots$ e da escolha do tipo de arredondamento. Em comum, as operações de arredondamento possuem a característica de produzir uma resposta da forma

$$\bar{x} := \pm(d_0{,}d_{-1}\ldots d_{-k+1})_b,$$

determinada "truncamento com k dígitos", ou da forma

$$\bar{x} \pm b^{-k+1} = \pm\left((d_0{,}d_{-1}\ldots d_{-k+1})_b + (0{,}0\ldots 1)_b\right),$$

que consiste em adicionar uma unidade ao dígito d_{-k+1} na representação $\bar{x}$.

A seguir estão enumerados os cinco tipos de arredondamento mais comuns. Os dois primeiros são arredondamentos para representante mais próximo, os demais são denominados "arredon-

damentos dirigidos".

1. Arredondamento para o mais próximo, desempate par: é o número com k dígitos mais próximo de x. Em caso de empate, isto é, se os dígitos sobressalentes $(0{,}d_{-k}d_{-k-1}\ldots)_b$ representarem a fração $\frac{1}{2}$, o dígito d_{-k+1} é arredondado para o algarismo seguinte se d_{-k+1} for ímpar e mantido se par. É a escolha padrão na maioria das situações.

   Regra:

$$\tilde{x} = \begin{cases} \bar{x}, & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots < 2^{-1} \quad \text{ou} \\ & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots = 2^{-1} \text{ e } d_{-k+1} \text{ for par;} \\ \\ \bar{x} \pm b^{-k+1}, & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots > 2^{-1} \quad \text{ou} \\ & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots = 2^{-1} \text{ e } d_{-k+1} \text{ for ímpar.} \end{cases}$$

2. Arredondamento para o mais próximo, desempate no sentido oposto ao zero: é o número com k dígitos mais próximo de x. Em caso de empate, isto é, se os dígitos sobressalentes $(0{,}d_{-k}d_{-k-1}\ldots)_b$ representarem a fração $\frac{1}{2}$, o dígito d_{-k+1} é arredondado para o algarismo seguinte.

   Regra:

$$\tilde{x} = \begin{cases} \bar{x}, & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots < 2^{-1} \\ \\ \bar{x} \pm b^{-k+1}, & \text{se } (0{,}d_{-k}d_{-k-1})_b\ldots \geq 2^{-1} \quad \text{ou} \end{cases}$$

3. Arredondamento por truncamento ou no sentido do zero: os dígitos sobressalentes são descartados. Possui esse nome pois o seu efeito é aproximar os números do zero.

   Regra:

$$\tilde{x} = \overline{x}.$$

4. Arredondamento no sentido de $+\infty$: o dígito d_{-k+1} é arredondado para o algarismo seguinte se o numeral for positivo e os dígitos sobressalentes forem não nulos. Se o numeral for negativo e os dígito sobressalentes forem não nulos, o dígito d_{-k} é mantido. Possui esse nome pois o seu efeito é deslocar os números de um valor maior ou igual a zero (ou seja no sentido de $+\infty$).

   Regra:

$$\tilde{x} = \begin{cases} \bar{x}, & \text{se } d_{-k} = d_{-k-1} = \ldots = 0 \quad \text{ou} \\ & \text{se } d_{-j} \neq 0 \text{ para algum } j \geq k \text{ e } x < 0; \\ \\ \bar{x} \pm b^{-k+1}, & \text{se } d_{-j} \neq 0 \text{ para algum } j \geq k \text{ e } x > 0. \end{cases}$$

5. Arredondamento no sentido de $-\infty$: o penúltimo dígito é mantido se o numeral for positivo e os dígitos sobressalentes forem não nulos. Se o numeral for negativo e os dígitos sobressalentes forem não nulos, o dígito d_{-k} é arredondado para o algarismo seguinte. Possui esse

nome pois o seu efeito é deslocar os números de um valor menor ou igual a zero (ou seja, no sentido de $-\infty$).

Regra:

$$\tilde{x} = \begin{cases} \bar{x}, & \text{se } d_{-k} = d_{-k-1} = \ldots = 0 \quad \text{ou} \\ & \text{se } d_{-j} \neq 0 \text{ para algum } j \geq k \text{ e } x > 0; \\ \bar{x} \pm b^{-k+1}, & \text{se } d_{-j} \neq 0 \text{ para algum } j \geq k \text{ e } x < 0. \end{cases}$$

Exemplo 3: O quadro abaixo apresenta o resultado das operações de arredondamento para três dígitos.

x	mais próximo, desempate par	mais próximo, desempate $\leftarrow 0 \rightarrow$	truncamento	sentido $+\infty$	sentido $-\infty$
13.140	13.100	13.100	13.100	13.200	13.100
0,01875	0,0188	0,0188	0,0187	0,0188	0,0187
-3,3196	-3,32	-3,32	-3,31	-3,31	-3,32
-2.145	-2.140	-2.150	-2.140	-2140	-2.150

1.2.4 Aritmética de ponto flutuante

As operações aritméticas com pontos flutuantes, simbolizadas pelos termos $\oplus$, $\ominus$, $\otimes$ e $\oslash$, e definidas sobre elementos de um mesmo sistema de ponto flutuante, retornam sempre um elemento desse mesmo sistema. Idealmente, as operações são realizadas como se os potnos flutuantes fossem números reais e então é utilizada uma operação de arredondamento para que o resultado seja um elemento do sistema de ponto flutuante.

Um outra propriedade importante dos pontos flutuantes diz respeito às propriedades algébricas que ao contrário dos reais, racionais e inteiros, em geral não são válidas. Vamos representar as operações de adição, subtração, multiplicação e divisão em ponto flutuante, respectivamente, pelos símbolos $\oplus$, $\ominus$, $\otimes$, $\oslash$. Dados três numeros com representação em ponto flutuante x, y e z, em geral

$$\begin{aligned} x \oplus y &\neq x + y, \\ x \otimes y &\neq x \times y, \\ (x \oplus y) \oplus z &\neq x \oplus (y \oplus z), \\ x \otimes (y \oplus z) &\neq (x \otimes y) \oplus (x \otimes z). \end{aligned}$$

Exemplo 4: Sejam $x = 1 \cdot 10^{-3}$ e $y = z = 1 \cdot 10^{10}$, elementos do conjunto $F(10, 3, -10, 10)$. Então

$$\begin{aligned} x \oplus (y \ominus z) &= 1{,}00 \cdot 10^{-3} \oplus (1{,}00 \cdot 10^{10} \ominus 1{,}00 \cdot 10^{10}) \\ &= 1{,}00 \cdot 10^{-3} \oplus 0 \\ &= 1{,}00 \cdot 10^{-3}, \end{aligned}$$

por outro lado

$$\begin{aligned}(x \oplus y) \ominus z &= (1{,}00 \cdot 10^{-3} \oplus 1{,}00 \cdot 10^{10}) \ominus 1{,}00 \cdot 10^{10} \\ &= 1{,}00 \cdot 10^{10} \ominus 1{,}00 \cdot 10^{10} \\ &= 0.\end{aligned}$$

1.2.5 Cancelamento catastrófico

É um efeito presente nas operações em ponto flutuante, caracterizado pelo aumento significativo do erro relativo no resultado da operação. O cancelamento catastrófico pode ser verificado principalmente na operação de subtração de dois pontos flutuantes muito próximos.

Exemplo 5: Vamos considerar a operação em ponto flutuante associada à subtração dos números racionais 0,98765432104234567 89 e 0,9876543209. Se os registros forem de 10 dígitos, a representação dos dois números será respectivamente

$$9{,}876543210 \times 10^{-1} \qquad \text{e} \qquad 9{,}876543209 \times 10^{-1}.$$

A diferença exata entre os dois números é de

$$1{,}423456789 \times 10^{-10}$$

enquanto que o resultado da operação de diferença em ponto flutuante é

$$1{,}00000000 \times 10^{-10},$$

uam diferença de aproximadamente 42%.

Um outro exemplo clássico é o das raízes de uma equação polinomial de segundo grau.

Exemplo 6: Seja a equação de segundo grau

$$x^2 + 400x - 0{,}00004617 = 0.$$

Essa equação possui duas raízes reais, uma próxima a -400 e outra próxima a 0. As raízes são dadas exatamente por

$$\frac{-400 - \sqrt{400^2 + 4 \times 0{,}00004617}}{2} \qquad \text{e} \qquad \frac{-400 + \sqrt{400^2 + 4 \times 0{,}00004617}}{2}.$$

A sequência de operações em ponto flutuante utilizada para calcular as raízes é dada por

$$(-b \ominus \text{sqrt}\,((b \otimes b) \ominus (4{,}000000000 \otimes c))) \oslash 2{,}000000000$$

e

$$(-b \oplus \text{sqrt}\,((b \otimes b) \ominus (4{,}000000000 \otimes c))) \oslash 2{,}000000000,$$

onde $b = 4{,}000000000 \times 10^2$ e $c = -4{,}617000000 \times 10^{-5}$. Substituindo os valores e realizando as operações, obtemos

$$
\begin{aligned}
& (-b \ominus \text{sqrt}\,((b \otimes b) \ominus (4{,}000000000 \otimes c))) \oslash 2{,}000000000 \\
= \;& \big(-4{,}000000000 \times 10^2 \\
& \ominus \text{sqrt}\left(1{,}600000000 \times 10^5 \oplus 1{,}846800000 \times 10^{-4}\right)\big) \oslash 2{,}000000000 \times 10^0 \\
= \;& \left(-4{,}000000000 \times 10^2 \ominus \text{sqrt}\left(1{,}600000002 \times 10^5\right)\right) \oslash 2{,}000000000 \times 10^0 \\
= \;& \left(-4{,}000000000 \times 10^2 \ominus 4{,}000000002 \times 10^2\right) \oslash 2{,}000000000 \times 10^0 \\
= \;& -4{,}000000001 \times 10^2,
\end{aligned}
$$

Realizando as operações para a outra raiz, obtemos o valor $1{,}000000000 \times 10^{-7}$. O valor da primeira raiz com os dezesseis primeiros dígitos exatos é $-400{,}0000001154249\ldots$ e há concordância com os dez primeiros dígitos obtidos na operação em ponto flutuante. O mesmo não ocorre com a segunda raiz. Neste caso, o valor com dezesseis dígitos exatos é $1{,}154249999666926\ldots \times 10^{-7}$ e, à exceção do primeiro dígito, todos os seguintes diferem o que caracteriza o cancelamento catastrófico.

A inexatidão no cálculo da segunda raiz pode ser diminuída consideravelmente se manipularmos a expressão de maneira a evitar a subtração de dois pontos flutuantes muito próximos. Analisando a expressão para a segunda raiz, podemos verificar que a operação inexata é a subtração presente na soma dos termos $-b$ e $\sqrt{b^2 - 4c}$. Neste exemplo, em valores absolutos, b é muito maior que c, portanto a representação em ponto flutuante do termo $\sqrt{b^2 - 4c}$ partilha muitos dígitos em comum com a representação de b.

Evitamos o cancelamento catastrófico em ponto flutuante realizando o cancelamento na própria expressão, antes de realizarmos as operações em ponto flutuante[19]:

$$
\begin{aligned}
-b + \sqrt{b^2 - 4c} \;& = \\
& = -b + |b| \sqrt{1 - \frac{4c}{b^2}} \\
& = -b + |b| \left(1 - \frac{2c}{b^2} + \ldots\right) \\
& \approx -b + |b| \left(1 - \frac{2c}{b^2}\right)
\end{aligned}
$$

[19] Na passagem da segunda para a terceira linha, foi realizada um expansão em série de potências para o termo $\sqrt{1-x}$ em torno de $x = 0$. Ou seja, levamos em consideração que o termo $\frac{4c}{b^2}$ é pequeno.

Como $b > 0$, a segunda raiz pode ser calculada a partir da aproximação[20]

$$\frac{-b + \sqrt{b^2 - 4c}}{2} \approx -\frac{c}{b}.$$

Substituindo os valores das constantes b, c e realizando as operações em ponto flutuante, obtemos a aproximação $1{,}154250000 \times 10^{-7}$ que possui um erro muitas vezes menor.

1.2.6 Padrão IEEE754

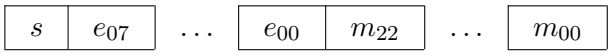
O padrão IEEE754 (a sigla se refere ao Institute of Electrical and Electronics Engineers) foi desenvolvido com o objetivo de unificar as diversas implementações em máquina de registros e operações em ponto flutuante. A maioria dos processadores atuais possuem unidades especializadas, denominadas FPUs (Unidades de Ponto Flutuante)[21] que suportam o padrão IEEE754 ou pelo menos suportam um subconjunto obrigatório das definições[22].

Além dos numerais em ponto flutuante, o padrão prevê que o registro pode conter informação sobre $+\infty$, $-\infty$, $+0$, -0, numerais subnormais (menores do que o usualmente suportado em uma notação $F(2, n, e_2, e_1)$) e os NaN ("not a number" reservado para operações ilegais como raízes de números negativos).

O padrão prevê quatro tipos de registros: registros de 32 bits denomindos pontos flutuantes de precisão simples, registros de 43 ou mais bits para precisão simples estendida, registros de 64 bits para precisão dupla e registros de 79 ou mais bits para precisão dupla estendida. A implementação de precisão simples é obrigatória, as demais são opcionais.

A título de ilustração, vale a pena estudar os registros de precisão simples.

Os 32 bits do registro de um numeral em precisão simples são divididos de acordo com o diagrama a abaixo,



o bit s é responsável pelo sinal, os bits $e_{07}e_{06}\ldots e_{00}$ representam o expoente e finalmente os bits $m_{22}m_{21}\ldots m_{00}$ representam o significando. Com os 8 bits do expoente representam-se inteiros entre 0 e 255, no entanto os registros relativos ao 0 (00000000) e 255 (11111111) são reservados para uso especial, sobram portanto os inteiros entre 1 e 254. Segundo o padrão, o inteiro relativo ao expoente está deslocado de 127, então os 8 bits permitem representar os valores inteiros entre -126 e 127. Os 23 bits restantes são utilizados para representar o significando com 24 dígitos binários (já que o primeiro dígito é sempre igual a 1 em uma base binária, não há necessidade explícita de armazená-lo no registro e com isso ganha-se um bit extra) com uma diferença: no padrão IEEE754 os pontos flutuantes normalizados começam com o primeiro dígito à esquerda

[20] Obtemos uma aproximação mais exata se levarmos em consideração os termos de ordem superior na expansão em série de potências. Porém, como os pontos flutuantes deste exemplo armazenam apenas dez dígitos, a inclusão dos termos adicionais não altera o resultado da operação (o leitor pode verificar esse fato).

[21] Em língua inglesa, *Floating Point Units*.

[22] Ao contrário das ALUs que também estão presentes nos circuitos de calculadoras científicas, o uso de FPUs é mais comum em processadores para computador. Por exemplo, os modelos de processadores da Intel anteriores ao 486DX não possuiam uma unidade FPU própria (apesar de ser possível a instalação de uma FPU independente no computador).

posicionado antes da vírgula. Portanto o registro de 32 bits é capaz de armazenar os elementos não nulos do sistema $F(2, 24, -126, 127)$ e mais os casos especiais:

1. zeros: bits do expoente e do significando todos nulos. O bit de sinal pode ser igual a 0 ou 1, ou seja, há uma representação para $+0$ e -0.
2. subnormais: bits do expoente todos nulos e os do significando e sinal guardam informação sobre o subnormal.
3. infinitos: bits do expoente todos iguais a 1 e os do significando iguais a 0. O bit de sinal pode ser igual a 0 ou 1, ou seja, há uma representação para $+\infty$ e $-\infty$.
4. NaN: bits do expoente todos iguais a 1 e os demais bits contém informação de diagnóstico.

Exemplo 7: Vamos encontrar o registro equivalente ao numeral 1345,875. O primeiro passo é representar o número na base binária:

$$1345{,}875 = 10101000001{,}111_2.$$

Em seguida vamos reescrevê-lo como um ponto flutuante normalizado: $1{,}0101000001111 \cdot 2^{10}$. Ignoramos o 1 antes da vírgula e adicionamos tantos 0 à direita quantos forem necessários para preencher os 23 bits, dessa forma encontramos os bits do significando:

$$01010000011110000000000.$$

O expoente vale 10, com o deslocamento de 127, o inteiro a ser representado pelos bits do expoente é o $137 = 10001001_2$. O bit de sinal é igual a 0 pois o número é positivo. O registro de 32 bits completo é dado por

$$01000100101010000011110000000000.$$

1.3 Erros

O principal propósito da computação científica é a construção de métodos que permitam obter aproximações numéricas para um dado objeto cujo valor exato seja impossível ou muito difícil de ser obtido. É fundamental que esse métodos produzam as aproximações do modo mais eficiente e acurado possível(muitas vezes é necessário chegar a um balanço aceitável entre essas duas propriedades). Tal objetivo sóp pode ser alcançado se as diferentes fontes de erro forem controladas.

1.3.1 Origem dos erros

Os resultados obtidos através de métodos numéricos podem ser afetados por muitos tipos de erros. Enquanto alguns deles podem ser difíceis de serem controlados, outros podem ser mitigados ou mesmo eliminados através de uma conveniente modificação do método.

De maneira geral, é possível classificar os erros que afetam o resultado de um procedimento computacional como erros nos dados de entrada, erros de arredondamento e erros de truncamento.

Os erros nos dados de entrada são aqueles relacionados à alguma medida física. Os aparelhos utilizados para medição sempre possuem uma precisão finita – em geral não muito grande quando comparada à precisão que é possível ser obtida na representação de números em máquinas– e nem sempre é possível melhorá-la consideravelmente.

Os erros de truncamento são os mais comuns em algoritmos numéricos. Ocorrem quando, de alguma maneira, é necessário aproximar um procedimento formado por uma sequência infinita de passos através de um outro procedimento finito.

Os erros de arredondamento são aqueles relacionados às limitações que existem na forma de representar números em máquinas.

Qualquer que seja a natureza do erro, o método numérico utilizado em um dado problema deveria ser capaz de estimar as suas consequências sobre o resultado: os dados de saída. Vamos discutir nesta seção como "propagar" e controlar essa incerteza presente nos valores nas diversas operações próprias a um procedimento computacional. Devido ao seu caráter, os erros de truncamento serão estudados em conjunto com os algoritmos que os geram.

1.3.2 Conceitos iniciais

Seja x um número que se conhece exatamente e $\tilde{x}$ uma representação finita, ou aproximação, de x, por exemplo $x = \pi$ e $\tilde{x} = 3{,}14159$, ou ainda $x = \dfrac{1}{3}$ e $\tilde{x} = 0{,}333$. Então definimos erro absoluto e erro relativo como:

Definição 1.3.1 (Erro absoluto e erro relativo). *O erro absoluto na representação $\tilde{x}$ é definido por $|x - \tilde{x}|$. O erro relativo é definido como $\dfrac{|x - \tilde{x}|}{|x|}$.*

Em uma máquina, denominamos "precisão" o número de dígitos no significando. Por outro lado, como o próprio nome diz, a "exatidão" ou "acurácia" de uma aproximação é uma medida de quanto ela está próxima do valor exato. Uma maneira de estimar a exatidão utiliza o conceito de dígitos exatos de uma representação.

Proposição 1.3.2 (Dígitos exatos de $\tilde{x}$)
Seja a aproximação $\tilde{x}$ de um número x em base b. O número de dígitos exatos em $\tilde{x}$ é um número natural k que satisfaz as desigualdades

$$-\log_b\left(\left|\frac{x-\tilde{x}}{x}\right|\right) - 1 < k < -\log_b\left(\left|\frac{x-\tilde{x}}{x}\right|\right) + 1$$

Através dessas definições podemos apreciar a diferença entre os conceitos de precisão e exatidão de uma representação $\tilde{x}$.

Demonstração: Seja x um número[23] representado pelo numeral $0{,}d_{-1}d_{-2}\ldots \times b^e$ e sua aproximação $\tilde{x}$, representada por um ponto flutuante normalizado com n dígitos de precisão dos quais apenas os k iniciais são iguais aos de x: $0{,}d_{-1}d_{-2}\ldots d_{-k}\tilde{d}_{-k-1}\ldots\tilde{d}_n \times b^e$, ou seja, necessariamente $d_{-k-1} \neq \tilde{d}_{-k-1}$.

Podemos verificar que x e $\tilde{x}$ podem ser também representados por

$$x = 0{,}d_{-1}d_{-2}\ldots d_{-k} \times b^e + 0{,}d_{-k-1}d_{-k-2}\ldots \times b^{e-k}$$

[23] Sem perda de generalidade, consideramos x e $\tilde{x}$ positivos.

e

$$\tilde{x} = 0{,}d_{-1}d_{-2}\ldots d_{-k} \times b^e + 0{,}\tilde{d}_{-k-1}\tilde{d}_{-k-2}\ldots\tilde{d}_{-n} \times b^{e-k}.$$

O menor valor absoluto para a diferença[24] entre eles é dado pela situação em que d_{-k-1} e $\tilde{d}_{-k-1}$diferem de uma unidade, todos os demais dígitos até o n-ésimo são iguais e $d_j = 0$ para $j > n$. Portanto

$$|x - \tilde{x}| \geq 0{,}1_b \times b^{e-k} = b^{e-k-1}.$$

Por outro lado, o maior valor absoluto para é dado pela situação limite na qual todos os dígitos $\tilde{d}_{-k-1}$,$\tilde{d}_{-k-2}$, $\ldots\ \tilde{d}_n = 0$ e todos os dígitos d_{-k-1},d_{-k-2}, $\ldots = b-1$. Portanto

$$\begin{aligned} |x - \tilde{x}| &< 0{,}(b-1)(b-1)\ldots_b \times b^{e-k} \\ &= \left((b-1)b^{-1} + (b-1)b^{-2} + \ldots\right) b^{e-k} \\ &= \left(1 - \frac{1}{b} + \frac{1}{b} - \frac{1}{b^2} + \ldots\right) b^{e-k} = b^{e-k}. \end{aligned}$$

Essas estimativa nos permite estabelecer limites inferiores e superiores para o erro relativo:

$$\frac{b^{e-k-1}}{|0{,}d_{-1}d_{-2}\ldots \times b^e|} \leq \frac{|x - \tilde{x}|}{|x|} < \frac{b^{e-k}}{|0{,}d_{-1}d_{-2}\ldots \times b^e|}.$$

Levando em conta que $|0{,}d_{-1}d_{-2}\ldots \times b^e| \geq 0{,}1_b \times b^e = b^{e-1}$ e que $|0{,}d_{-1}d_{-2}\ldots \times b^e| < 0{,}(b-1)(b-1)\ldots_b \times b^e = b^e$, temos finalmente que

$$b^{-k-1} < \left|\frac{x - \tilde{x}}{x}\right| < b^{-k+1}.$$

Como todas as quantias na expressão acima são positivas e a função logaritmo é monótona :

$$-k - 1 < \log_b \left|\frac{x - \tilde{x}}{x}\right| < -k + 1$$

■

Exemplo 8: Seja $x = \pi$ e $\tilde{x} = 3{,}141675 = 0{,}3141675 \cdot 10^1$. Nesse caso a precisão da representação $\tilde{x}$ é de 7 dígitos, no entanto $\tilde{x}$ possui apenas 4 dígitos exatos. Se conhecêssemos apenas o erro relativo $\frac{|x - \tilde{x}|}{x} \approx 0{,}0000262117$, de acordo com a proposição, teríamos que o número de dígitos exatos é um natural k tal que $3{,}58151... < k < 5{,}58151...$, ou seja, poderíamos concluir que $\tilde{x}$ possui entre 4 e 5 dígitos exatos.

1.3.3 Propagação de erros

Vamos assumir que $\tilde{x}$ é uma aproximação de um número x, sem perda de generalidade, vamos supor que $x > \tilde{x}$. Se quisermos encontrar o valor da função f calculada em x mas só dispormos

[24] Aqui a operação de diferença é a usual para os números reais.

de $\tilde{x}$ devemos aproximar $f(x)$ por $f(\tilde{x})$. Se a função f for diferenciável, pelo teorema do valor médio (veja o apêndice, teorema 10.8.1) existe um $\varepsilon \in (\tilde{x}, x)$:

$$f(x) - f(\tilde{x}) = f'(\varepsilon)(x - \tilde{x}).$$

Porém o teorema não diz nada sobre ε além de sua existência. Para estimar o erro cometido na aproximação $f(\tilde{x})$ devemos utilizar mais informações sobre a função f. Por exemplo, se soubermos que a derivada de f é limitada no intervalo $(\tilde{x}, x)$, $\sup_{y\in(\tilde{x},x)} |f'(y)| = M$, então temos que

$$|f(x) - f(\tilde{x})| = |f'(\varepsilon)|\ |x - \tilde{x}| \leq M|x - \tilde{x}|.$$

Mesmo que não sejamos capazes de determinar o valor máximo da derivada da função f no intervalo $(\tilde{x}, x)$, ainda podemos considerar o caso em que $|\tilde{x} - x| := \delta\tilde{x}$ é pequeno. Nesse caso, se $f'(\tilde{x}) \neq 0$ e f' não varia muito rapidamente na vizinhança de $\tilde{x}$ então como $\delta\tilde{x}$ é muito pequeno podemos considerar $M \approx f'(\tilde{x})$ na expressão acima, dessa forma

$$\delta f(\tilde{x}) := |f(x) - f(\tilde{x})| \approx f'(\tilde{x})|x - \tilde{x}| = f'(\tilde{x})\delta\tilde{x}.$$

A generalização dessa fórmula no caso em que f depende de mais de uma variável é obtida através de uma parametrização de f como uma função de uma única variável.

Para tanto, vamos considerar um conjunto aberto conexo $U \subseteq \mathbb{R}^n$ no qual $f : U \to \mathbb{R}$ é uma função contínua e diferenciável. A partir de dois elementos $x, \tilde{x} \in U$ tais que a combinação convexa $(1 - s)\tilde{x} + sx$ pertence a U para qualquer $s \in (0, 1)$, segue da diferenciabiliade de f que a função $g : (0, 1) \to \mathbb{R}$, definida por $g(s) := f\left((1 - s)\tilde{x} + sx\right)$, é contínua e diferenciável no intervalo aberto $(0, 1)$. O teorema do valor médio garante a existência de um $\varepsilon \in (0, 1)$ tal que $g(1) - g(0) = g'(\varepsilon)$ e portanto a existência de um $\bar{x} \in U$, $\bar{x} = (1 - \varepsilon)\tilde{x} + \varepsilon x$ tal que

$$\begin{aligned} f(x) - f(\tilde{x}) &= \nabla f(\bar{x}) \cdot (x - \tilde{x}) \\ &= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\bar{x})\,(x_j - \tilde{x}_j)\,. \end{aligned}$$

Em valor absoluto, a desigualdade triangular implica

$$\begin{aligned} |f(x) - f(\tilde{x})| &= \left|\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\bar{x})\,(x_j - \tilde{x}_j)\right| \\ &\leq \sum_{j=1}^{n} \left|\frac{\partial f}{\partial x_j}(\bar{x})\,(x_j - \tilde{x}_j)\right| \\ &= \sum_{j=1}^{n} \left|\frac{\partial f}{\partial x_j}(\bar{x})\right| |x_j - \tilde{x}_j| \end{aligned}$$

e seguindo o mesmo roteiro utilizado na função de uma única variável chegamos à fórmula

$$\delta f(\tilde{x}) \approx \sum_{j=1}^{n} \left|\frac{\partial f}{\partial x_j}(\tilde{x})\right| \delta x_i.$$

Um procedimento semelhante pode ser utilizado no caso de funções que dependam de duas ou mais variáveis. Seja o erro na i-ésima variável $\delta\tilde{x_i}$ de um função f que dependa de n variáveis, então o erro propagado na função f resultado dos erros nas n variáveis é dado por $\delta f(\tilde{x_1}, \tilde{x_2}, \ldots, \tilde{x_n})$:

$$\delta f(\tilde{x_1}, \tilde{x_2}, \ldots, \tilde{x_n}) \approx \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\tilde{x_1}, \tilde{x_2}, \ldots, \tilde{x_n}) \right| \delta\tilde{x_i}.$$

1.3.4 Instabilidade numérica

Alguns problemas matemáticos e algoritmos numéricos possuem a propriedade de ampliar drasticamente os erros presentes nos dados de entrada e assim invalidar a saída ou resposta. No contexto do cálculo numérico, esse fenômeno é denominado *instabilidade numérica*. A instabilidade numérica pode estar relacionada às propriedades matemáticas do problema ou então à estrutura do algoritmo utilizado para resolvê-lo. De qualquer maneira, ao estudar um problema que pretendemos reescrever numericamente é imprescindível a análise de estabilidade do algoritmo ou o *condicionamento do problema matemático*.

Dizemos que um problema é *mal condicionado* quando pequenas variações nos dados de entrada resultam em grandes variações na resposta. O seguinte exemplo é ilustrativo.

Exemplo 9: Seja o polinômio[25] $P(x) = (x-1)(x-2)\ldots(x-20) = x^{10} - 210\,x^9 + \ldots + 20!$. Ou seja, P é um polinômio em x com raízes inteiras $1, 2, 3, \ldots, 20$.

Vamos considerar agora o polinômio $\tilde{P}$:

$$\tilde{P}(x) = P(x) + 2^{-23}\,x^{19}.$$

Ou seja, $\tilde{P}$ é igual P a menos de um erro relativo de $5{,}7 \cdot 10^{-10}$ no coeficiente do termo x^9. As raízes do polinômio $\tilde{P}$ são (com cinco dígitos)

$x_1 = 1{,}0000\ldots$	$x_6 = 5{,}9999\ldots$	$x_{10} = 10{,}8929\ldots + 1{,}1493\ldots\,i$	$x_{11} = \overline{x_{10}}$
$x_2 = 2{,}0000\ldots$	$x_7 = 7{,}0003\ldots$	$x_{12} = 12{,}8217\ldots + 2{,}1234\ldots\,i$	$x_{13} = \overline{x_{12}}$
$x_3 = 3{,}0000\ldots$	$x_8 = 7{,}9930\ldots$	$x_{14} = 15{,}3059\ldots + 2{,}7753\ldots\,i$	$x_{15} = \overline{x_{14}}$
$x_4 = 4{,}0000\ldots$	$x_9 = 9{,}1472\ldots$	$x_{16} = 18{,}1813\ldots + 2{,}5489\ldots\,i$	$x_{17} = \overline{x_{16}}$
$x_5 = 5{,}0000\ldots$	$x_{20} = 9{,}5020\ldots$	$x_{18} = 20{,}4767\ldots + 1{,}0390\ldots\,i$	$x_{19} = \overline{x_{18}}$

Um erro relativo de $5{,}7 \cdot 10^{-10}$ no coeficiente x^{19} foi capaz de alterar drasticamente parte das raízes. Como entender esse fenômeno?

Vamos avaliar quanto varia cada raiz x_j ($j = 1, 2, \ldots, 20$) quando alteramos o coeficiente -210 do termo x^{19}. Para tanto vamos considerar o novo polinômio $\mathfrak{P}(\alpha, x) = P(x) - \alpha x^{19}$. Esse novo polinômio possui vinte raízes $x_j(\alpha)$, $j = 1, 2, \ldots, 20$, cada uma delas depende de α. A derivada $\left.\dfrac{dx_j}{d\alpha}\right|_{\alpha=0}$ contém informação sobre quanto a i-ésima raíz de $\mathfrak{P}(\alpha, x)$ varia com relação a α quando $\alpha = 0$. Como $\mathfrak{P}(0, x) \equiv P(x)$, essas derivadas possuem a informação

[25] Esse polinômio é conhecido como um dos polinômios de Wilkinson. Foi introduzido em 1963 por James H. Wilkinson para ilustrar o mal-condicionamento na determinação de raízes de polinômios. Veja: Wilkinson, J.H. "The perfidious polynomial". em Golub, G. H. "Studies in Numerical Analysis". Mathematical Association of America. pg. 3. (1984).

que buscamos sobre as raízes de P.

Para calcular $\left.\frac{dx_j}{d\alpha}\right|_{\alpha=0}$ não será necessário encontrar explicitamente a dependência das raízes em α, basta lembrar que por definição $\mathfrak{P}(\alpha, x_j(\alpha)) = 0$ ($x_j(\alpha)$ é raiz do polinômio $\mathfrak{P}(\alpha, x)$) e portanto de acordo com a regra da cadeia

$$\frac{d}{d\alpha}(\mathfrak{P}(\alpha, x_j(\alpha))) = \frac{\partial \mathfrak{P}}{\partial \alpha}(\alpha, x_j(\alpha)) + \frac{\partial \mathfrak{P}}{\partial x}(\alpha, x_j(\alpha))\frac{dx_j}{d\alpha} = 0$$

e assim,

$$\frac{dx_j}{d\alpha} = -\frac{\frac{\partial \mathfrak{P}}{\partial \alpha}(\alpha, x_j(\alpha))}{\frac{\partial \mathfrak{P}}{\partial x}(\alpha, x_j(\alpha))}.$$

O numerador da expressão acima é simplesmente $\frac{\partial \mathfrak{P}}{\partial \alpha}(\alpha, x_j(\alpha)) = -x^{19}$. Para calcular $\frac{\partial \mathfrak{P}}{\partial x}(\alpha, x_j(\alpha))$, basta notar que $\mathfrak{P}(\alpha, x) = (x-1)(x-2)\ldots(x-20) - \alpha x^{19}$, portanto $\frac{\partial \mathfrak{P}}{\partial x}(\alpha, x) = (x-2)\ldots(x-20)+(x-1)(x-3)\ldots(x-20)+\ldots+(x-1)\ldots(x-19)-19\alpha x^{18}$, ou seja

$$\frac{\partial \mathfrak{P}}{\partial x}(\alpha, x) = -19\alpha x^{18} + \sum_{k=1}^{20} \prod_{l=1, l\neq k}^{20} (x-l).$$

Assim,

$$\frac{dx_j}{d\alpha} = -\frac{-x_j^{19}(\alpha)}{-19\alpha x^{18} + \sum_{k=1}^{20} \prod_{l=1,l\neq k}^{20}(x_j(\alpha) - l)}.$$

Quando $\alpha = 0$, $\mathfrak{P}(0, x) = P(x)$ então $x_j(0) = j$ para $j = 1, 2, \ldots, 20$. Ou seja,

$$\left.\frac{dx_j}{d\alpha}\right|_{\alpha=0} = -\frac{-j^{19}}{\sum_{k=1}^{20} \prod_{l=1,l\neq k}^{20}(j-l)} = -\frac{-j^{19}}{\prod_{l=1,l\neq j}^{20}(j-l)}.$$

Na expressão anterior, o somatório foi retirado pois o único termo que contribuía era o $k = j$ (Por que?).

Por economia de notação, vamos denominar a derivada da j-ésima raiz em $\alpha = 0$ por ζ_j:

$$\left.\frac{dx_j}{d\alpha}\right|_{\alpha=0} \equiv \zeta_j = -\frac{-j^{19}}{\prod_{l=1,l\neq j}^{20}(j-l)}. \tag{1.3.1}$$

Para determinar o valor de ζ_j basta utilizar a expressão (1.3.1):

$\zeta_1 \approx 8{,}2 \cdot 10^{-18}$	$\zeta_6 \approx -5{,}8 \cdot 10^{1}$	$\zeta_{11} \approx 4{,}6 \cdot 10^{7}$	$\zeta_{16} \approx -2{,}4 \cdot 10^{9}$
$\zeta_2 \approx -8{,}2 \cdot 10^{-11}$	$\zeta_7 \approx 2{,}5 \cdot 10^{3}$	$\zeta_{12} \approx -2{,}0 \cdot 10^{8}$	$\zeta_{17} \approx 1{,}9 \cdot 10^{9}$
$\zeta_3 \approx 1{,}6 \cdot 10^{-6}$	$\zeta_8 \approx -6{,}0 \cdot 10^{4}$	$\zeta_{13} \approx 6{,}1 \cdot 10^{8}$	$\zeta_{18} \approx -1{,}0 \cdot 10^{9}$
$\zeta_4 \approx -2{,}1 \cdot 10^{-3}$	$\zeta_9 \approx 8{,}4 \cdot 10^{5}$	$\zeta_{14} \approx -1{,}3 \cdot 10^{9}$	$\zeta_{19} \approx 3{,}1 \cdot 10^{8}$
$\zeta_5 \approx 6{,}1 \cdot 10^{-1}$	$\zeta_{10} \approx -7{,}6 \cdot 10^{6}$	$\zeta_{15} \approx 2{,}1 \cdot 10^{9}$	$\zeta_{20} \approx -4{,}3 \cdot 10^{7}$

E assim podemos notar que enquanto a variação no valor das primeiras raízes é pequena, o mesmo não pode ser dito para as demais. O valor das derivadas ilustra o mal condicionamento deste problema.

O próximo exemplo contempla a instabilidade numérica devido ao algoritmo.

Exemplo 10: Seja o seguinte algoritmo para calcular a integral $I_n := \int_0^1 dx\, x^n\, e^{x-1}$, onde $n = 1, 2, \ldots$ Podemos integrar por partes I_n:

$$\begin{aligned} I_n &= x^n\, e^{x-1}\big|_{x=0}^{x=1} - n\int_0^1 dx\, x^{n-1}\, e^{x-1} \\ &= 1 - n\, I_{n-1}. \end{aligned} \tag{1.3.2}$$

Como $I_n = 1 - nI_{n-1}$, se conhecermos I_1, poderemos encontrar I_2, I_3 e assim por diante. Antes de dar prosseguimento vamos examinar um pouco mais a integral I_n. No intervalo de integração $[0, 1]$, o integrando é sempre positivo, além disso, o termo e^{x-1} é sempre menor ou igual a 1, isto implica a desigualdade

$$0 < I_n \leq \int_0^1 dx\, x^n = \frac{1}{n+1}. \tag{1.3.3}$$

Portanto, devemos esperar que os termos I_n decresçam com n e sejam sempre positivos. O primeiro termo I_1 pode ser calculado explicitamente utilizando integração por partes,

$$I_1 = \int_0^1 dx\, x\, e^{x-1} = 1 - \int_0^1 dx\, e^{x-1} = e^{-1}.$$

Iremos representar I_1 aproximadamente por $\tilde{I}_1$ com cinco casas após a vírgula:

$$\tilde{I}_1 = 0{,}367879.$$

Nesse caso $\delta = |I_1 - \tilde{I}_1| \approx 5{,}6 \cdot 10^{-7}$. Se calcularmos os demais termos a partir da iteração (1.3.2) encontraremos

$$\begin{array}{ll} \tilde{I}_1 = 0{,}367879 & \tilde{I}_6 = 0{,}127120 \\ \tilde{I}_2 = 0{,}264242 & \tilde{I}_7 = 0{,}110160 \\ \tilde{I}_3 = 0{,}207274 & \tilde{I}_8 = 0{,}118720 \ \ (?) \\ \tilde{I}_4 = 0{,}170904 & \tilde{I}_9 = -0{,}0684800 \ \ (??) \\ \tilde{I}_5 = 0{,}145480 & \tilde{I}_{10} = 1{,}68480 \ \ (???) \end{array}$$

O que ocorreu aqui? Ao contrário do observado nas últimas igualdades acima, a estimativa (1.3.3) prevê o decrescimento de I_n com n e $I_n > 0$. Vamos analisar a iteração levando em conta a aproximação inicial. Partimos portanto de $\tilde{I}_1 = I_1 + \delta$, então segundo a iteração (1.3.2)

$$\tilde{I}_2 = 1 - 2(\tilde{I}_1) = 1 - 2(I_1 + \delta) = I_2 - 2\delta$$

e então

$$\tilde{I}_3 = 1 - 3(\tilde{I}_2) = 1 - 3(I_2 - 2\delta) = I_3 + 3 \cdot 2 \cdot \delta.$$

Por indução temos então que

$$\tilde{I}_n = I_n + (-1)^{n+1} n!\delta.$$

Ou seja, o erro cresce com o fatorial de n. Dessa forma

$$\tilde{I_9} \approx I_9 + 9!\,5{,}6 \cdot 10^{-7} \approx I_9 + 0,2$$

e

$$\tilde{I_{10}} \approx I_{10} - 10!\,5{,}6 \cdot 10^{-7} \approx I_{10} - 2.$$

Felizmente uma pequena alteração do algoritmo permite resolver a questão da instabilidade. Basta reescrever a iteração (1.3.2) como

$$I_{n-1} = \frac{1}{n}(1 - I_n) \tag{1.3.4}$$

isolando I_{n-1}. Utilizando os mesmos critérios é possível verificar que o erro cometido no ponto inicial é mais e mais diluído a cada passo do algoritmo (verifique!)

Além dessa propriedade, a diferença da iteração (1.3.4) reside no fato de que o ponto inicial deve ser um valor de n suficientemente grande para que os índices abaixo possam ser calculados pelo algoritmo. Agora a estimativa (1.3.3) é útil, pois ela nos diz que I_n decresce com n e satisfaz a desigualdade $I_n < \dfrac{1}{n+1}$. Portanto podemos utilizar um número n suficientemente alto e considerar a aproximação $\tilde{I}_n = 0$ ou $\tilde{I}_n = \dfrac{1}{n+1}$.

1.4 Exercícios

1) Represente em base decimal os seguintes numerais:

a) $1011101{,}101_2$

b) $7{,}7_8$

c) $AB{,}FE_{16}$

2) Utilize no máximo dez algarismos após a vírgula para representar o número 0,11 nas bases 2, 3 e 16.

3) Quantos bits, no mínimo, são necessários para representar os inteiros entre -3000 e 3000 em um registro de máquina. Quais os inteiros adicionais que podem ser representados utilizando o mesmo registro?

4) Dado o número $7+\frac{1}{4}+\frac{1}{16}$, determine duas bases (menores ou iguais a 9) nas quais o número pode ser representado em notação posicional com um número finito de dígitos. Represente o número nessas bases.

5) Quantas soluções não negativas admite a equação $1 \oplus x = 1$ em uma máquina que utilize o sistema de ponto flutuante $F(10, 5, -99, 99)$.

6) Represente todos os elementos positivos do conjunto $F(2, 3, -1, 2)$ na reta real.

7) Qual das seguintes maneiras de calcular o polinômio $x^2 - 1$ é a mais exata no sistema $F(10, 6, -20, 20)$ quando $x = 1{,}0001$?

- $(x \otimes x) \ominus 1$
- $(x \oplus 1) \otimes (x \ominus 1)$
- $(x \otimes (x \ominus 1)) \oplus (x \ominus 1)$

8) O seguinte registro hexadecimal corresponde a um ponto flutuante de 16 bits no padrão IEEE754: $\boxed{\text{ABCD}}$. O bits mais significativo representa o sinal, os 5 bits seguintes subtraídos de 15 representam o expoente e os demais o significando. Determine o número representado em base decimal. Supondo o arredondamento par, qual é o menor ponto flutuante nesse sistema que adicionado a 1 retorna um ponto flutuante maior do que 1?

9) Realize a operação de adição dos números $1{,}50390625 \times 10^{-1}$ e $5{,}8203126 \times 10^{-1}$ no sistema de ponto flutuante de base 4, com 4 dígitos e arredondamento par. A resposta deve estar em base 10.

10) Calcule o valor da expressão $\sqrt{\cos\left(\frac{\pi}{2} - x\right)}$ para $x = 0{,}490000 \times 10^{-4}$ em um sistema de ponto flutuante com 6 dígitos de precisão (serão 3 operações). Agora leve em consideração que para x pequeno e positivo $\cos\left(\frac{\pi}{2} - x\right) \approx x$ e recalcule o valor da expressão. Se assumirmos que a segunda forma é a mais acurada, qual o erro relativo cometido no cálculo da primeira expressão?

11) A frequência natural de ressonância de um circuito RLC é dada pela expressão

$$f_0 = \frac{1}{2\pi}\frac{1}{\sqrt{LC}}.$$

Determine uma expressão para estimativa de erro no cálculo de f_0 se $C = 10\mu F \pm 10\%$ e $L = 2{,}5 \times 10^{-1}\mu H \pm 5\%$.

12) Como devemos reescrever a expressão seguinte

$$2 + x - \sqrt{x^2 + 4}$$

para valores pequenos de x no sistema $F(10, 10, -20, 20)$ de modo que o erro de arredondamento cometido seja pequeno. Experimente com os valores $x = \{1{,}000000 \cdot 10^{-6}; 1{,}000000 \cdot 10^{-7}; 1{,}000000 \cdot 10^{-8}\}$.

13) Dados dois pontos flutuantes x e y, determine em quais situações devemos reescrever a expressão

$$e^{\frac{x^2}{2}} - \sqrt{e^{x^2} + \cos y}$$

para que a mesma possa ser estimada sem grandes perdas de exatidão.

14) Determine a representação do número 20,25 em um ponto flutuante normalizado em base 3 com 8 dígitos.

15) Considere a função de duas variáveis $f(x, y) = 2 + \cos(x)\cos(y)$. Determine uma estimativa para o erro relativo cometido ao calcularmos f em $x = y = \pi/4$ se o erro absoluto nas variáveis for 10^{-6}.

16) Um registro de ponto flutuante com 10 bits possui a seguinte estrutura: o 1º bit guarda informação sobre o sinal, os três bits seguintes guarda informação sobre o expoente (deslocado de três unidades) e os seis restantes guardam os dígitos do significando (a partir do segundo dígito, pois o primeiro é sempre igual a 1). Assim, o registro 1110001000 representa o número $(-1)^1 \times 0{,}1001000_2 \times 2^{6-3}$, ou seja, representa $-(\frac{1}{2} + \frac{1}{16}) \times 8 = -4{,}5$. Qual seria o registro que representa o número 6,125?

17) A inversa da função cosseno hiperbólico $\cosh(x) := \dfrac{e^x + e^{-x}}{2}$ é denominada arco-cosseno hiperbólico $\cosh^{-1}(x) = -\ln\left(x - \sqrt{x^2 - 1}\right)$. Reescreva a expressão para $\cosh^{-1}$ de modo que o seu valor possa ser calculado sem cancelamentos catastróficos. Determine uma aproximação com oito dígitos para $\cosh^{-1}(7{,}47764 \times 10^{57})$ a partir dessa nova expressão.

18) Utilize a fórmula para a propagação de erros e obtenha uma estimativa para o comportamento do erro relativo cometido na função $f(x, y) := y\sqrt{1 + x^2}$ em termos dos erros relativos cometidos em x e y.

19) Reescreva as expressões

$$\sqrt{e^{2x} + 1} \; - \; e^x \qquad \text{e} \qquad \sqrt{e^{2x} + x^2} \; - \; e^x$$

de modo que seja possível obter os seus valores para $x = 100$ utilizando a aritmética de ponto flutuante da máquina ("doubles" no Scilab).

20) Determine a representação do número 20,25 em um ponto flutuante normalizado em base 3 com 8 dígitos.

21) Considere uma máquina cujo registro de ponto flutuante dispõe de espaço para a representação de 10 dígitos em base 3 para o significando. Determine uma aproximação com três dígitos para o erro relativo cometido na representação do número 0,1 nesse registro.

22) Trabalhe no base $F(10, 4, -10, 10)$ com arredondamento dado pela operação de truncamento e a partir da expressão

$$\ln(1+x) \approx x \otimes (1 \ominus (0{,}5 \otimes (x \otimes (1 \ominus (0{,}6666 \otimes x)))))$$

determine uma aproximação para $\ln(0{,}8)$. Quanto vale o erro relativo (em módulo)?

23) Um setor de de disco de ângulo θ e raio r possui área $A(r, \theta) = \frac{1}{2}\theta r^2$. Determine o erro relativo na medida da área de um setor de raio 2 e ângulo $\frac{\pi}{4}$ se o erro relativo na medida do raio é 0,01 e no ângulo é 0,01.

24) Reescreva a função:

$$f(x) = \ln\left(10^{16} + x\right) - \ln\left(10^{16} - x\right),$$

de modo que não ocorram grandes perdas de exatidão quando x for um número pequeno. Determine uma aproximação para $x = 13$ no sistema $F(10, 5, -20, 20)$ a partir da nova expressão para f. (Sugestão: coloque o termo dominante em evidência e utilize o fato de que x é pequeno e portanto $\ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3}$).

25) Um registro de ponto flutuante com 10 bits possui a seguinte estrutura: o 1º bit guarda informação sobre o sinal, os três bits seguintes guarda informação sobre o expoente (deslocado de três unidades) e os seis restantes guardam os dígitos do significando (a partir do segundo dígito, pois o primeiro é sempre igual a 1). Assim, o registro 1110001000 representa o número $(-1)^1 \times 0{,}1001000_2 \times 2^{6-3}$, ou seja, representa $-(\frac{1}{2}+\frac{1}{16})\times 8 = -4{,}5$. Qual seria o registro que representa o número 6,125?

26) Considere um registro de ponto flutuante que dispõe de 12 bits para o signifcando. Qual é o erro relativo associado à representação do número $\left(110{,}\overline{101}\right)_2 = 6 + \frac{5}{7}$ nesse registro? (Obs.1: lembre que não há necessidade de guardar o 1º dígito no registro. Obs.2.: leve em consideração o arredondamento par.)

2 Sistemas de equações lineares

Os sistemas de equações lineares fazem parte da descrição matemática dos mais diversos fenômenos em todas as áreas das ciências naturais e também são peça fundamental de diversos algoritmos utilizados em computação. Por exemplo, mais adiante na disciplina, veremos que através da discretização dos domínios onde está definida uma equação diferencial é possível reduzi-la a um sistema de equações lineares. Em outras aplicações como mecânica dos fluidos e mecânica estrutural, não é incomum trabalhar com sistemas de ordem 10^5ou mais.

Devido a sua quase onipresença , é muito importante poder contar com técnicas que encontre a solução para os sistemas de modo eficiente e acurado. Atualmente podemos contar com software de alta qualidade para resolver sistemas de equações lineares e ainda hoje este assunto está sendo ativamente pesquisado. Principalmente a resolução de sistemas muito grandes em "clusters" de computadores e computadores com processadores vetoriais. Para melhor compreender e utilizar esses softwares é essencial conhecer as propriedades dos algoritmos mais simples.

Vamos nos concentrar no estudo de sistemas de n equações e n incógnitas $x_1, x_2, \ldots, x_n$:

$$\begin{cases} a_{1,1}x_1 & + & a_{1,2}x_2 & + & \ldots & + & a_{1,n}x_n & = & b_1 \\ a_{2,1}x_1 & + & a_{2,2}x_2 & + & \ldots & + & a_{2,n}x_n & = & b_2 \\ & & & & \vdots & & & & \\ a_{n,1}x_1 & + & a_{n,2}x_2 & + & \ldots & + & a_{n,n}x_n & = & b_n \end{cases} \tag{2.0.1}$$

onde os coeficientes a_{ij} e as constantes b_i são números reais.

Denominamos o conjunto de todas as possíveis soluções de um sistema linear de *conjunto solução*. Dados dois sistemas lineares, dizemos que os dois são *equivalentes* se possuírem o mesmo conjunto solução.

Sob um ponto de vista geométrico, consideramos cada variável x_i nas equações como a i-ésima componente de um ponto no $\mathbb{R}^n$. Dessa forma, uma vez escolhida uma variável x_j, cada equação do sistema relaciona x_j às demais variáveis. Ou seja, cada equação representa um subespaço de dimensão $n-1$ contido no $\mathbb{R}^n$. O conjunto solução é formado pela intersecção de todos esses subespaços.

O conjunto solução de um sistema linear pode ser vazio, conter um único elemento (uma única solução) ou infinitos elementos (infinitas soluções). Por exemplo, no caso de um sistema de duas variáveis e duas incógnitas, cada equação representa um reta no $\mathbb{R}^2$ e o conjunto solução será vazio se as retas forem paralelas, possuirá apenas uma solução se as retas se cruzarem ou ainda infinitas soluções se as retas se sobreporem.

O sistema de equações (2.0.1) é convenientemente representado de forma matricial através da equação

$$\mathrm{A} \cdot \mathbf{x} = \mathbf{b}, \tag{2.0.2}$$

onde

$$\mathrm{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{e} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Nessa forma, a matriz A e o vetor coluna **b** constituem os dados de entrada do problema numérico. A solução, ou dado de saída, é dado pelo vetor coluna **x** cujas componentes serão aproximadas através de métodos numéricos.

Um representação alternativa para o sistema (2.0.1) consiste na *matriz completa do sistema* formada pela justaposição das matrizes A e **b** da equação (2.0.2):

$$[\,\mathrm{A}|\,\mathbf{b}] := \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} & b_n \end{pmatrix}. \tag{2.0.3}$$

Essa representação é útil pois concatena todos os dados de entrada e permite descrever o sistema de equações (2.0.1) do modo mais econômico.

Uma das propriedades mais evidentes do sistema de equações (2.0.1) é o fato de que sua solução independe da ordem em que as equações são postas. Ou seja, se o vetor $\mathbf{x_0}$ é solução de (2.0.2), ele também será solução de um sistema em que as equações são as mesmas em uma ordem diferente. Portanto ao trocarmos duas linhas quaisquer da matriz completa (2.0.3) obtemos uma matriz equivalente[1]. A operação de troca de linhas na matriz completa é um tipo de operação conhecido como *operação elementar*.

As operações elementares são operações realizadas em uma matriz completa que resultam em uma matriz equivalente. Por ser linear, o sistema de equações (2.0.1) possui a seguinte propriedade: a troca de qualquer equação pela combinação linear dela com qualquer outra equação do sistema resulta em um sistema com o mesmo conjunto solução. Portanto, a substituição de qualquer linha de (2.0.3) pela combinação linear da linha a ser substituída com qualquer outra, não altera a solução. Essa operação é portanto uma operação elementar. Note que a simples multiplicação de uma linha por uma constante não nula é um caso especial dessa operação.

A partir das operação elementares é possível implementar uma série de métodos distintos para encontrar a solução do sistema. Os métodos que utilizam uma sequência finita de operações elementares para determinar a solução são denominados *métodos diretos*.

[1] Dadas duas matrizes completas C e D, dizemos que são equivalentes se representam sistemas de equações equivalentes. Notação:

$$\mathrm{C} \sim \mathrm{D}.$$

2.1 Métodos diretos

Um dos métodos diretos mais conhecidos é a eliminação de Gauss-Jordan. Esse método consiste em aplicar sucessivas operações elementares até que a matriz completa $[\mathrm{A}|\,\mathbf{b}]$ assuma a forma

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 & x_1^* \\ 0 & 1 & 0 & \dots & 0 & x_2^* \\ 0 & 0 & 1 & & 0 & x_3^* \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & x_n^* \end{pmatrix} \sim [\mathrm{A}|\,\mathbf{b}]\,. \tag{2.1.1}$$

Esse é o método comumente utilizado na solução de pequenos sistemas de n incógnitas e n equações na disciplina "álgebra linear":

1. Partimos da primeira coluna à esquerda em $[\mathrm{A}|\,\mathbf{b}]$ e, sucessivamente, utilizamos operações elementares para anular as componentes abaixo da diagonal e igualar as componentes da diagonal a 1 até chegar à última linha. Nesse ponto teremos uma matriz triangular superior.

2. Então retornamos a partir da última linha e da direita para a esquerda, realizamos operações elementares para anular as componentes acima da diagonal anulando elementos.

Se o sistema possuir uma única solução, o resultado do método será a matriz completa reduzida no lado esquerdo da relação de equivalência (2.1.1). Os termos $x_1^*, \dots, x_n^*$ são a solução do sistema.

Exemplo 11: Seja o sistema de equações lineares

$$\begin{cases} x_1 - 2x_2 + x_3 = 0 \\ \quad\;\; 2x_2 - 8x_3 = 8 \\ -4x_1 + 5x_2 + 9x_3 = -9 \end{cases}$$

representado pela matriz completa

$$\mathrm{A_c} = \begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ -4 & 5 & 9 & -9 \end{pmatrix}.$$

É comum utilizar a seguinte notação

$$\mathrm{A_c} = \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix} \sim \begin{pmatrix} L_1 \\ 3L_1 - 2L_2 \\ L_3 \end{pmatrix} = \begin{pmatrix} L_1 \\ \tilde{L_2} \\ L_3 \end{pmatrix} \sim \begin{pmatrix} L_3 \\ \tilde{L_2} \\ L_1 \end{pmatrix}$$

para representar operações elementares nas linhas de uma matriz. De acordo com essa nota-

ção

$$
\begin{aligned}
\mathrm{A_c} &\sim \begin{pmatrix} L_1 \\ L_2 \\ 4L_1 + L_3 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ 0 & -3 & 13 & -9 \end{pmatrix} = \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix} \\
&\sim \begin{pmatrix} L_1 \\ \frac{1}{2}L_2 \\ \frac{3}{2}L_2 + L_3 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -4 & 4 \\ 0 & 0 & 1 & 3 \end{pmatrix} = \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix} \\
&\sim \begin{pmatrix} L_1 - L_2 \\ L_2 + 4L_3 \\ L_3 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 0 & -3 \\ 0 & 1 & 0 & 16 \\ 0 & 0 & 1 & 3 \end{pmatrix} = \\
&\sim \begin{pmatrix} L_1 + 2L_2 \\ L_2 \\ L_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -29 \\ 0 & 1 & 0 & 16 \\ 0 & 0 & 1 & 3 \end{pmatrix}. \qquad (2.1.2)
\end{aligned}
$$

Portanto a matriz $\mathrm{A_c}$ é equivalente a matriz (2.1.2) que representa o sistema de equações lineares

$$
\begin{cases} x_1 = 29 \\ x_2 = 16 \\ x_3 = 3 \end{cases},
$$

ou seja, a solução do sistema é $(x_1, x_2, x_3)^T = (29, 16, 3)^T$.

Existem outros métodos diretos que consistem na decomposição da matriz A em produtos de matrizes, como a decomposição LU, a decomposição de Cholesky (quando A é simétrica) e a decomposição singular (quando A não é quadrada). Outras classes de métodos diretos são construídos quando a matriz A possui uma estrutura especial, como no caso das matrizes esparsas. Não os estudaremos aqui[2] mas sim uma alternativa ao método Gauss-Jordan conhecida como eliminação gaussiana.

2.1.1 Eliminação Gaussiana

Uma alternativa ao método de Gauss-Jordan consiste em parar a sequência de operações elementares quando a matriz completa for da forma

$$
[\mathrm{U}|\,\mathbf{c}] \sim [\mathrm{A}|\,\mathbf{b}]\,,
$$

[2] Existe muita literatura sobre esses tópicos, sugerimos ao leitor mais interessado o texto:

- Stoer, J. ; Bulirsch, R. *An introduction to Numerical Analysis*, Springer-Verlag, 1983.

onde U é triangular superior:

$$\mathrm{U} = \begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \dots & u_{1,n} \\ 0 & u_{2,2} & u_{2,3} & \dots & u_{2,n} \\ 0 & 0 & u_{3,3} & \dots & u_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_{n,n} \end{pmatrix}.$$

A matriz completa $[\mathrm{U}|\,\mathbf{c}]$ representa o sistema de equações

$$\begin{cases} u_{1,1}x_1 + u_{1,2}x_2 + \dots + u_{1,n}x_n = c_1 \\ u_{2,2}x_2 + \dots + u_{2,n}x_n = c_2 \\ \vdots \\ u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n = c_{n-1} \\ u_{n,n}x_n = c_n \end{cases}$$

que pode ser resolvido através de substituições recursivas: a partir da última linha, temos que

$$x_n = \frac{c_n}{u_{n,n}}. \tag{2.1.3}$$

Utilizamos o valor de x_n obtido em (2.1.3) para determinar x_{n-1} através da equação imediatamente superior,

$$x_{n-1} = \frac{1}{u_{n-1,n-1}}\left(c_{n-1} - u_{n-1,n}x_n\right).$$

Em seguida, utilizamos os valores de x_n e x_{n-1} para determinar x_{n-2}; e assim, sucessivamente, teremos para a i-ésima incógnita x_i:

$$x_i = \frac{1}{u_{i,i}}\left(c_i - \sum_{j=i+1}^{n} u_{i,j}x_j\right),$$

onde $i = 1, 2, \dots, n-1$.

2.1.2 Estabilidade do método

Se houvesse uma máquina capaz de armazenar variáveis com precisão arbitrária e realizasse operações aritméticas sem erros de arredondamento em tempo finito, através do método de eliminação gaussiana descrito na subseção anterior, seriamos perfeitamente capazes de encontrar a solução exata para um sistema de equações lineares. No entanto, devemos nos recordar que na prática as variáveis são representadas por pontos flutuantes e consequentemente, as operações aritméticas estão sujeitas a erros de arredondamento.

Para ilustrar as limitações da eliminação gaussiana simples vamos considerar o seguinte exemplo.

Exemplo 12: Seja um sistema de equações lineares representado pela matriz completa $[\mathrm{A}|\,\mathbf{b}]$. Vamos supor que estamos na fase final de um processo de eliminação gaussiana e que as ope-

rações envolvem pontos flutuantes. Vamos supor também que, com exceção de uma particular componente, o resultado de todas as operações representam exatamente a matriz completa equivalente na forma (quase) escalonada:

$$\begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,n-2} & u_{1,n-1} & u_{1,n} & c_1 \\ 0 & u_{2,2} & u_{2,3} & \cdots & u_{2,n-2} & u_{2,n-1} & u_{2,n} & c_2 \\ 0 & 0 & u_{3,3} & \cdots & u_{3,n-2} & u_{3,n-1} & u_{3,n} & c_3 \\ \vdots & \vdots & & \ddots & & \vdots & \vdots & \vdots \\ 0 & 0 & & \cdots & 0 & \epsilon & 1 & 1 \\ 0 & 0 & & \cdots & 0 & 2 & 1 & 3 \end{pmatrix}_{n\times(n+1)} \nsim [\,\mathrm{A}|\,\mathbf{b}\,]\,. \tag{2.1.4}$$

Se as operações fossem realizadas sem erros de arredondamento, ao invés da matriz (2.1.4) teríamos o escalonamento exato dado por

$$\begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,n-2} & u_{1,n-1} & u_{1,n} & c_1 \\ 0 & u_{2,2} & u_{2,3} & \cdots & u_{2,n-2} & u_{2,n-1} & u_{2,n} & c_2 \\ 0 & 0 & u_{3,3} & \cdots & u_{3,n-2} & u_{3,n-1} & u_{3,n} & c_3 \\ \vdots & \vdots & & \ddots & & \vdots & \vdots & \vdots \\ 0 & 0 & & \cdots & 0 & 0 & 1 & 1 \\ 0 & 0 & & \cdots & 0 & 2 & 1 & 3 \end{pmatrix}_{n\times(n+1)} \sim [\,\mathrm{A}|\,\mathbf{b}\,]\,. \tag{2.1.5}$$

e as duas últimas componentes da solução exata são dadas por

$$x_n = 1 \qquad \text{e} \qquad x_{n-1} = 1.$$

Por outro lado, o valor obtido para as duas últimas componentes da solução aproximada a partir da matriz completa (2.1.4) são resultado das operações

$$\tilde{x}_n = \frac{3 - \dfrac{2}{\epsilon}}{1 - \dfrac{2}{\epsilon}} \qquad \text{e} \qquad \tilde{x}_{n-1} = \frac{1 - \tilde{x}_n}{\epsilon}. \tag{2.1.6}$$

Vamos supor que as variáveis pertençam ao sistema de ponto flutuante $F(10, 10, -90, 90)$ e que o erro seja pequeno, por exemplo, $\epsilon = 10^{-11}$. Nesse caso, $\dfrac{2}{\epsilon} = 2 \times 10^{11}$ e assim o numerador e denominador de $\tilde{x}_n$, $3 - \dfrac{2}{\epsilon}$ e $1 - \dfrac{2}{\epsilon}$ serão arredondados respectivamente para $-1{,}999999999 \times 10^{11}$ e $-1{,}999999999 \times 10^{11}$ ou $2{,}000000000 \times 10^{11}$ e $2{,}000000000 \times 10^{11}$, dependendo da escolha de arredondamento. Em qualquer um dos casos, para um erro $\epsilon = 10^{-11}$ teremos como soluções aproximadas

$$\tilde{x}_n = 1{,}000000000 \times 10^{0} \qquad \text{e} \qquad \tilde{x}_{n-1} = 0.$$

Uma análise mais atenta dos termos (2.1.6) nos permite concluir que nesse exemplo não é possível encontrar a solução exata através da eliminação gaussiana simples.

Essa dificuldade em encontrar a solução exata apesar de um erro pequeno e isolado poderia nos levar a pensar que a eliminação gaussiana seria inútil na busca da solução através de métodos numéricos. Este não é o caso. O problema explicitado no exemplo foi drasticamente amplificado pelo fato de que na posição de pivô da $(n-1)$-ésima coluna da matriz completa a componente possui um valor muito pequeno ϵ. De acordo com o algoritmo do método de eliminação gaussiana simples, o resultado é a combinação linear das duas últimas linhas com um termo multiplicativo de valor absoluto muito grande $-\frac{2}{\epsilon}$. Uma medida simples como a operação elementar de troca das duas últimas linhas antes da combinação linear das mesmas resultaria nas aproximações

$$\tilde{x}_n = \frac{1-\frac{3}{2}\epsilon}{1-\frac{1}{2}\epsilon} \qquad \text{e} \qquad \tilde{x}_{n-1} = \frac{3-\tilde{x}_n}{2}$$

que conduzem à solução exata quando o erro ϵ for suficientemente pequeno[3].

Essa observação induz a um aprimoramento da eliminação gaussiana denominada *método de eliminação gaussiana com pivoteamento parcial* (EGPP) que consiste na troca de linhas em uma coluna pivô para que a posição de pivô seja ocupada pela maior componente em valor absoluto antes de realizarmos as operações que eliminam os termos abaixo da posição de pivô na mesma coluna. Esse método atenua o tipo de problema ilustrado no exemplo, no entanto é possível verificar que a escolha da componente de maior valor absoluto não elimina o problema por completo nem mesmo é sempre a melhor escolha possível para a posição de pivô.

Em geral, associa-se a estabilidade do método EGPP à demonstração de que a solução calculada através do método corresponde à solução exata do sistema original adicionado de uma "pequena perturbação" que consiste em termos adicionados aos coeficientes da matriz original. Essa questão é tratada por James H. Wilkinson em um artigo clássico de 1961. Em valor absoluto, as perturbações possuem um crescimento que é limitado por uma expressão da forma $const. \times n^2 \times \rho_n$, onde n é a dimensão da matriz e ρ_n é denominado "fator de crescimento":

$$\rho_n := \frac{\max_{i,j,k} \left| a_{i,j}^{(k)} \right|}{\max_{i,j,k} \left| a_{i,j}^{(k)} \right|},$$

e $a_{i,j}^{(k)}$ é o valor do coeficiente da matriz no k-ésimo passo da EGPP.

Existem exemplos nos quais $\rho_n = 2^{n-1}$ mas em muitos dos sistemas associados à aplicações práticas, o fator de crescimento é limitado em n, ou seja, não cresce mais do que uma constante. Esse é o caso das matrizes totalmente não negativas (matrizes cujas submatrizes possuem determinantes não negativos), das matrizes simétricas positivas definidas e das diagonais dominantes (por linha ou coluna).

[3] Naturalmente, devemos levar em conta que as operações são em ponto flutuante, ou seja, a representação obtida é exata sem a necessidade de tomar o limite $\epsilon \to 0$, basta que seja um valor suficientemente próximo de 0.

2.1.3 Condicionamento em sistemas de equações lineares

Deve-se notar que mesmo quando existe estabilidade, isso não significa necessariamente que o erro cometido na aproximação seja pequeno. Se a matriz de coeficientes for "próxima" de uma matriz singular, a solução obtida é muito sensível a alterações nos dados.

Em vista dessas observações, naturalmente estaremos diante da questão de decidir entre duas ou mais aproximações aceitáveis qual delas é a mais adequada, ou "próxima" da solução exata. Porém nesse caso, o objeto que chamamos de solução não é um simplesmente um número real e sim um vetor do espaço vetorial n-dimensional $\mathbb{R}^n$. Como comparar esses objetos? Compararemos os vetores através de *normas*.

No contexto de um espaço vetorial no $\mathbb{R}^n$, a norma corresponde à noção intuitiva de distância. Assim, como o valor absoluto da diferença entre dois reais x e y, $|x-y|$, pode ser entendida como a "distância" entre eles[4], a norma irá desempenhar um papel semelhante para elementos do espaço vetorial $\mathbb{R}^n$. No entanto, há uma diferença importante, não existe uma única possibilidade de definição para norma.

Definição 2.1.1 (Norma para um espaço vetorial). *Dado um espaço vetorial V, uma norma em V, simbolizada por $\|\cdot\|$, é uma função $\|\cdot\| : V \to \mathbb{R}_+$ com as seguintes propriedades, para quaisquer $x, y \in V$ e $\alpha \in \mathbb{R}$(ou $\mathbb{C}$)*

1. $\|x\| \geq 0$ *e* $\|x\| = 0$ *se e somente se* $x = \mathbf{0}$*, o elemento nulo de* V*, (positividade).*
2. $\|\alpha x\| = |\alpha| \ \|x\|$*, (homogeneidade).*
3. $\|x + y\| \leq \|x\| + \|y\|$*, (desigualdade triangular).*

A seguinte proposição é um resultado imediato da definição de norma.

Proposição 2.1.2
Dados dois elementos x,y de um espaço vetorial V e uma norma qualquer $\|\cdot\|$, definida nesse espaço, a seguinte desigualdade é válida

$$\|x - y\| \geq |\|x\| - \|y\|| \,.$$

Demonstração: A demonstração segue da desigualdade triangular e da propriedade de homogeneidade.

$$\begin{aligned} \|x\| &= \|(x-y)+y\| \\ &\leq \|x-y\| + \|y\| \,, \end{aligned}$$

o que implica

$$\|x - y\| \geq \|x\| - \|y\| \,.$$

Levando em conta a homogeneidade da norma e a desiguladade anterior,

$$\|x - y\| = \|y - x\| \geq \|y\| - \|x\|$$

[4]Em particular, quando eles guardam informação sobre posição em um sistema de coordenadas, neste caso, $|x|$ corresponde à distância de x à origem.

o que demonstra o resultado. ■

Note que a definição nada diz sobre a natureza do espaço vetorial V nem fornece uma prescrição sobre como a norma é construída de fato. Basta determinar uma função que satisfaça os critérios da definição. No nosso caso, os sistemas de equações lineares $\mathrm{A}\mathbf{x} = \mathbf{b}$ envolvem apenas vetores $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ e matrizes $\mathrm{A} \in \mathbb{M}_{m\times n}$(conjunto de todas as matrizes de m linhas e n colunas com componentes reais). Os conjuntos dos vetores do $\mathbb{R}^n$, assim como os das matrizes $\mathbb{M}_{m\times n}$, munidos das operações usuais formam espaços vetoriais para quaisquer m e n naturais não nulos.

Exemplo 13 (Normas para o $\mathbb{R}^n$): Em geral, diferenciamos as normas entre si por alguma notação asicionada ao símbolo $\|\cdot\|$. Em particular, vamos considerar as normas $\|\cdot\|_1$, $\|\cdot\|_2$ e $\|\cdot\|_\infty$: dado um vetor $\mathbf{x} \in \mathbb{R}^n$, de componentes $(\mathbf{x})_i = x_i$, definimos as normas

1. $\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$, (norma 1),

2. $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$, (norma 2 ou norma euclideana),

3. $\|\mathbf{x}\|_\infty := \max_i |x_i|$, (norma sup ou norma ∞).

É um exercício simples verificar que essas três normas satisfazem as propriedades da definição.

Da mesma forma vamos considerar as normas $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_F$ e $\|\cdot\|_\infty$ para matrizes $\mathrm{A} \in \mathbb{M}_{m\times n}$:

Exemplo 14 (Normas para o $\mathbb{M}_{m\times n}$): Dada uma matriz $\mathrm{A} \in \mathbb{M}_{m\times n}$, de componentes $a_{i,j}$, definimos as normas

1. $\|\mathrm{A}\|_1 := \max_{1\le j\le n} \sum_{i=1}^m |a_{i,j}|$, (norma 1),

2. $\|\mathrm{A}\|_2 := \sqrt{\max \sigma\left(A^T A\right)}$, (norma 2),

3. $\|\mathrm{A}\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2} \ge \|\mathrm{A}\|_2$, (norma Frobenius),

4. $\|\mathrm{A}\|_\infty := \max_{1\le i\le m} \sum_{j=1}^n |a_{i,j}|$, (norma sup ou norma ∞),

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

É um exercício simples verificar que essas três normas satisfazem as propriedades da definição.

Utilizamos os mesmos índices para identificar algumas das normas de vetores e matrizes pois, nesse caso, elas possuem uma importante relação entre si, elas são normas *compatíveis*: dado qualquer vetor $\mathbf{x} \in \mathbb{R}^n$, qualquer matriz $\mathrm{A} \in \mathbb{M}_{m\times n}$ e o vetor que resulta da multiplicação matricial $\mathrm{A}\mathbf{x} \in \mathbb{R}^m$, verificamos que, de acordo com as definições de norma para vetores e matrizes dos exemplos anteriores,

$$\|\mathrm{A}\mathbf{x}\|_\alpha \le \|\mathrm{A}\|_\alpha \|\mathbf{x}\|_\alpha$$

nos casos em que $\alpha = 1, 2$ ou ∞. Ou melhor, essas normas para matrizes são construídas a partir da definição da norma vetorial compatível de acordo com a definição

$$\|\mathrm{A}\|_\alpha := \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathrm{A}\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha} = \max_{\mathbf{y}\in\mathbb{R}^n:\|\mathbf{y}\|_\alpha=1} \|\mathrm{A}\mathbf{y}\|_\alpha .$$

A partir das definições de norma para vetores e matrizes, é possível comparar diferentes aproximações para a solução de um sistema de equações lineares. Se $\mathbf{x}^0$ e $\mathbf{x}^1$ são duas aproximações distintas do sistema não singular (possui apenas uma única solução)$\mathrm{A}\mathbf{x} = \mathbf{b}$, então a multiplicação matricial de A pelos vetores que representam a aproximação resultam em duas aproximações para o vetor das constantes $\mathbf{b}$:

$$\mathrm{A}\mathbf{x}^0 = \mathbf{b}^0 \qquad \text{e} \qquad \mathrm{A}\mathbf{x}^1 = \mathbf{b}^1.$$

Parece razoável supor que, através dos resíduos $\left\|\mathbf{b} - \mathbf{b}^0\right\|$ e $\left\|\mathbf{b} - \mathbf{b}^1\right\|$, os vetores $\mathbf{b}^0$ e $\mathbf{b}^1$ fornecem uma boa indicação sobre a exatidão das aproximações $\mathbf{x}^0$ e $\mathbf{x}^1$. O exemplo a seguir mostra que isso não é necessariamente verdade.

Exemplo 15: Seja o sistema de equações lineares representado pela seguinte matriz completa

$$\begin{pmatrix} 0{,}780 & 0{,}563 & 0{,}217 \\ 0{,}913 & 0{,}659 & 0{,}254 \end{pmatrix}$$

e duas matrizes constituídas por pequenas perturbações da matriz $\mathbf{b}$:

$$\tilde{\mathbf{b}} = \mathbf{b} + \begin{pmatrix} -0{,}1343 \times 10^{-3} \\ -0{,}1572 \times 10^{-3} \end{pmatrix} \qquad \text{e} \qquad \tilde{\tilde{\mathbf{b}}} = \mathbf{b} + \begin{pmatrix} -0{,}1 \times 10^{-6} \\ 0 \end{pmatrix}.$$

A matriz formada pelas duas primeiras colunas da matriz completa não é singular, portanto o sistema possui uma única solução, dada por $\mathbf{x}^* = \begin{pmatrix} 1 & -1 \end{pmatrix}^T$.

Já as soluções dos sistemas $A\mathbf{x} = \tilde{\mathbf{b}}$ e $A\mathbf{x} = \tilde{\tilde{\mathbf{b}}}$ são dadas respectivamente por Os resíduos associados às aproximações são dados respectivamente por

$$\tilde{\mathbf{x}} = \begin{pmatrix} 0.999 \\ -1.001 \end{pmatrix} \qquad \text{e} \qquad \tilde{\tilde{\mathbf{x}}} = \begin{pmatrix} 0.341 \\ -0.087 \end{pmatrix}$$

Note que qualquer que seja a norma utilizada, $\left\|\mathbf{x}^* - \tilde{\tilde{\mathbf{x}}}\right\| > \|\mathbf{x}^* - \tilde{\mathbf{x}}\|$:

$$\begin{aligned} \left\|\mathbf{x}^* - \tilde{\tilde{\mathbf{x}}}\right\|_1 = 1{,}572 &> \|\mathbf{x}^* - \tilde{\mathbf{x}}\|_1 = 2 \times 10^{-3}, \\ \left\|\mathbf{x}^* - \tilde{\tilde{\mathbf{x}}}\right\|_2 = 1{,}125\ldots &> \|\mathbf{x}^* - \tilde{\mathbf{x}}\|_2 = 1{,}414\ldots \times 10^{-3}, \\ \left\|\mathbf{x}^* - \tilde{\tilde{\mathbf{x}}}\right\|_\infty = 0{,}913 &> \|\mathbf{x}^* - \tilde{\mathbf{x}}\|_\infty = 10^{-3}. \end{aligned}$$

Como o erro contido em $\tilde{\mathbf{b}}$ é maior do que o contido em $\tilde{\tilde{\mathbf{b}}}$, poderíamos ter sido levados a crer que a aproximação $\tilde{\tilde{\mathbf{x}}}$ é a mais acurada, no entanto podemos constatar facilmente que $\|\mathbf{x}^* - \tilde{\mathbf{x}}\|_\alpha < \left\|\mathbf{x}^* - \tilde{\tilde{\mathbf{x}}}\right\|_\alpha$ qualquer que seja a norma.

Esse comportamento está relacionado ao fato da matriz A ser próxima de uma matriz singular. Matrizes com tal característica são denominadas matrizes mal-condicionadas. O condicionamento de uma matriz contém informação sobre a degradação da solução de um sistema $A\mathbf{x} = \mathbf{b}$ quando o vetor $\mathbf{b}$ é perturbado pela adição de um termo $\delta\mathbf{b}$. De maneira mais precisa, dado que $\mathbf{x}$ é solução do sistema, o condicionamento de A é uma constante multiplicativa que estabelece um limite superior para o erro relativo na solução $\dfrac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}$, onde $\mathbf{x} + \delta\mathbf{x}$ é solução do sistema

$$A\left(\mathbf{x} + \delta\mathbf{x}\right) = \mathbf{b} + \delta\mathbf{b}. \tag{2.1.7}$$

O primeiro termo resultante da multipicação no lado esquerdo de (2.1.7) é cancelado pelo primeiro termo do lado direito de (2.1.7). Multiplicando a equação restante pela inversa de A teremos

$$\delta\mathbf{x} = \mathrm{A}^{-1}\delta\mathbf{b}.$$

Agora, a escolha de normas vetoriais e matriciais compatíveis aplicadas à equação anterior leva à desiguladade

$$\begin{aligned} \|\delta\mathbf{x}\|_\alpha &= \left\|\mathrm{A}^{-1}\delta\mathbf{b}\right\|_\alpha \\ &\leq \left\|\mathrm{A}^{-1}\right\|_\alpha \|\delta\mathbf{b}\|_\alpha, \end{aligned}$$

A partir da desigualdade anterior implica a seguinte expressão para o erro relativo na solução

$$\frac{\|\delta\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha} \leq \left\|\mathrm{A}^{-1}\right\|_\alpha \|\delta\mathbf{b}\|_\alpha \frac{1}{\|\mathbf{x}\|_\alpha}. \tag{2.1.8}$$

Para reescrever o lado direito da desigualdade acima como o erro relativo do vetor das constantes $\mathbf{b}$, basta utilizar o fato de que a sua norma pode ser limitada pelo produto das normas da matriz e do vetor solução :

$$\|\mathbf{b}\|_\alpha = \|\mathrm{A}\mathbf{x}\|_\alpha \leq \|\mathrm{A}\|_\alpha \|\mathbf{x}\|_\alpha,$$

ou seja

$$\frac{1}{\|\mathbf{x}\|_\alpha} \leq \frac{\|\mathrm{A}\|_\alpha}{\|\mathbf{b}\|_\alpha}. \tag{2.1.9}$$

Utilizando a desigualdade (2.1.9) no último termo do lado direito de (2.1.8) concluímos que

$$\frac{\|\delta\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha} \leq \left\|\mathrm{A}^{-1}\right\|_\alpha \|\mathrm{A}\|_\alpha \frac{\|\delta\mathbf{b}\|_\alpha}{\|\mathbf{b}\|_\alpha}. \tag{2.1.10}$$

A partir da desigualdade (2.1.10) definimos o condicionamento de uma matriz inversível (ou seja, não singular) A como o valor numérico $\kappa_\alpha(A) := \left\|\mathrm{A}^{-1}\right\|_\alpha \|\mathrm{A}\|_\alpha$. Não é difícil verificar que o condicionamento é um número maior ou igual a unidade: se $\mathrm{I} = \mathrm{A}\mathrm{A}^{-1}$ é a matriz identidade, $\mathrm{I}\mathbf{x} = \mathbf{x}$, então para um vetor $\mathbf{x} \neq \mathbf{0}$

$$\|\mathbf{x}\|_\alpha = \left\|\mathrm{A}\mathrm{A}^{-1}\mathbf{x}\right\|_\alpha \leq \|\mathrm{A}\|_\alpha \left\|\mathrm{A}^{-1}\mathbf{x}\right\|_\alpha \leq \|\mathrm{A}\|_\alpha \left\|\mathrm{A}^{-1}\right\|_\alpha \|\mathbf{x}\|_\alpha$$

o que implica $\kappa_\alpha(A) \geq 1$.

Então se o erro relativo no valor da norma α de $\mathbf{b}$ for de ordem 10^{-p_b}, podemos esperar um erro relativo na norma α da solução, limitado superiormente por 10^{-p_s} e relacionado à p_b através da desigualdade

$$p_s \geq p_b - \log_{10} \kappa_\alpha(A)$$

que resulta do logaritmo em base 10 da desigualdade (2.1.10).

Ainda assim, devemos lembrar que a desigualdade (2.1.10) fornece apenas uma estimativa superior para o erro relativo. Na seção seguinte estudaremos um método intermediário entre os métodos diretos e os métodos iterativos, trata-se do método de refinamento iterativo de uma solução.

O desenvolvimento descrito anteriormente leva em consideração que a matriz A é exata. No entanto, os mesmos motivos que nos obrigam a considerar um vetor perturbado $b + \delta b$ – por exemplo, erros experimentais e erros de arredondamento decorrentes da representação em ponto flutuante – também estão presentes na representação das componentes da matriz A. Por esse motivo, a análise do sistema perturbado

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \tag{2.1.11}$$

é também necessária.

O desenvolvimento da equação (2.1.11) levará a uma desigualdade semelhante à (2.1.10). Nela é possível notar a presença do condicionamento, assim como a importância da perturbação δA. Mas esse resultado depende do seguinte lema.

Lema 2.1.3
Seja N uma matriz quadrada com $\|N\| < 1$. Então $(I + N)$ é inversível e

$$\left\|(I + N)^{-1}\right\| \leq \frac{1}{1 - \|N\|}. \tag{2.1.12}$$

Demonstração: Dado qualquer $\mathbf{x} \neq \mathbf{0}$ de dimensões compatíveis as de A temos que

$$\|(I + N)\mathbf{x}\| = \|\mathbf{x} - (-N\mathbf{x})\| \geq \|\mathbf{x}\| - \|N\mathbf{x}\| \geq (1 - \|N\|)\|\mathbf{x}\|,$$

ou seja, se $\|N\| < 1$ então $\|(I + N)\mathbf{x}\| \geq 0$ para qualquer $\mathbf{x} \neq \mathbf{0}$, consequentemente $(I + N)\mathbf{x} = \mathbf{0}$ possui apenas solução trivial e $(I + N)$ é inversível.

Como $(I + N)$ é inversível podemos considerar

$$\begin{aligned}
\|\mathbf{x}\| &= \left\|(1 + N)(1 + N)^{-1}\mathbf{x}\right\| \\
&= \left\|(1 + N)^{-1}\mathbf{x} - (-N)(1 + N)^{-1}\mathbf{x}\right\| \\
&\geq \left\|(1 + N)^{-1}\mathbf{x}\right\| - \left\|N(1 + N)^{-1}\mathbf{x}\right\| \\
&\geq (1 - \|N\|)\left\|(1 + N)^{-1}\mathbf{x}\right\|
\end{aligned}$$

que implica a desigualdade (2.1.12). ■

A partir desse lema, segue o seguinte resultado para o sistema (2.1.11).

Teorema 2.1.4
Sejam A e δA matrizes $n \times n$, A não singular, $\mathbf{b} \neq \mathbf{0}$ e $\delta\mathbf{b}$ vetores do $\mathbb{R}^n$ (representados por matrizes $n \times 1$). Sejam também $\mathbf{x}$ e $\delta\mathbf{x}$, vetores do $\mathbb{R}^n$ que satisfazem as equações

$$A\mathbf{x} = \mathbf{b}$$

e

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}.$$

Se para um norma matricial compatível à norma vetorial utilizada para os vetores, for válida a desigualdade

$$\left\|A^{-1}\delta A\right\| < 1$$

então

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{1}{1 - \|A^{-1}\delta A\|}\left(\kappa(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \left\|A^{-1}\delta A\right\|\right).$$

Se além disso, também for válida a desiguladade

$$\kappa(A)\frac{\|\delta A\|}{\|A\|} < 1$$

então

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\dfrac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta A\|}{\|A\|}\right).$$

Demonstração: Inicialmente abrimos os produtos cancelamos os termos relativos ao sistema original (sem as perturbações δ) e isolamos o termo associado a $\delta\mathbf{x}$.

$$\begin{aligned}
(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) &= \mathbf{b} + \delta\mathbf{b} \\
A\delta\mathbf{x} + \delta A\mathbf{x} + \delta A\delta\mathbf{x} &= \delta\mathbf{b} \\
A\left(I + A^{-1}\delta A\right)\delta\mathbf{x} &= \delta\mathbf{b} - \delta A\mathbf{x} \\
\delta\mathbf{x} &= \left(I + A^{-1}\delta A\right)^{-1}A^{-1}\delta\mathbf{b} - \left(I + A^{-1}\delta A\right)^{-1}A^{-1}\delta A\mathbf{x}.
\end{aligned}$$

Em norma, seguem as desiguladades

$$\begin{aligned}
\|\delta\mathbf{x}\| &\leq \left\|\left(I + A^{-1}\delta A\right)^{-1}\right\| \left\|A^{-1}\right\| \|\delta\mathbf{b}\| + \left\|\left(I + A^{-1}\delta A\right)^{-1}\right\| \left\|A^{-1}\delta A\right\| \|\mathbf{x}\| \\
&\leq \frac{1}{1 - \|A^{-1}\delta A\|} \left(\left\|A^{-1}\right\| \|\delta\mathbf{b}\| + \left\|A^{-1}\delta A\right\| \|\mathbf{x}\|\right) \\
&= \frac{1}{1 - \|A^{-1}\delta A\|} \left(\left\|A^{-1}\right\| \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \|\mathbf{b}\| + \left\|A^{-1}\delta A\right\| \|\mathbf{x}\|\right) \\
&\leq \frac{1}{1 - \|A^{-1}\delta A\|} \left(\|A\| \left\|A^{-1}\right\| \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \|\mathbf{x}\| + \left\|A^{-1}\delta A\right\| \|\mathbf{x}\|\right),
\end{aligned}$$

onde a passagem da primeira para segunda linha é resultado do lema anterior, e da terceira para a quarta é resultado da desiguladade $\|\mathbf{b}\| \leq \|A\| \|\mathbf{x}\|$. Em termos do condicionamento, chegamos à desigualdade

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{1}{1 - \|A^{-1}\delta A\|} \left(\kappa(A) \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \left\|A^{-1}\delta A\right\|\right).$$

Por outro lado, como

$$\left\|A^{-1}\delta A\right\| \leq \left\|A^{-1}\right\| \|\delta A\| = \left\|A^{-1}\right\| \|A\| \frac{\|\delta A\|}{\|A\|} = \kappa(A) \frac{\|\delta A\|}{\|A\|},$$

se $\kappa(A) \dfrac{\|\delta A\|}{\|A\|} < 1$, então

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \dfrac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta A\|}{\|A\|}\right).$$

■

As desigualdades do teorema anterior contém o resultado (2.1.10).

Essas estimativas permitem avaliar o impacto das alterações nos dados de entrada na solução do problema. No entanto, todas estão relacionadas ao condicionamento da matriz de coeficientes cuja avaliação envolve o conhecimento da inversa da matriz. Uma alternativa é a estimativa proposta por Prager e Oettli[5] que pode ser colocada da seguinte forma. Seja $\tilde{x}$ a solução aproximada de um sistema $Ax = b$, que no entanto é a solução de um outro sistema $\tilde{A}\tilde{x} = \tilde{b}$. O teorema desenvolvido por esses autores permite relacionar as matrizes desses sistemas ao resíduo da solução aproximada $\tilde{x}$.

Teorema 2.1.5 (Prager e Oettli, 1964)
Seja $\tilde{x}$ uma solução aproximada do sistema $Ax = b$. Então existem matrizes $\tilde{A}$, $\tilde{b}$, Δ_A e Δ_b que satisfazem $\tilde{A}\tilde{x} = \tilde{b}$ e estão sujeitas às desigualdades $\left|\left(A - \tilde{A}\right)_{i,j}\right| \leq (\Delta_A)_{i,j}$ e $\left|b_i - \tilde{b}_i\right| \leq (\Delta_b)_i$,

[5]Referência:

- Oettli,W; Prager,W. *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides.* Numer. Math. **6** (1964) 405–409.

se e somente se para cada componente i,

$$|(r(\tilde{x}))_i| \leq \sum_j (\Delta_A)_{i,j} |\tilde{x}_j| + (\Delta_b)_i \tag{2.1.13}$$

onde $r(\tilde{x}) := b - A\tilde{x}$ é o resíduo.

Demonstração: Vamos supor que existem matrizes $\tilde{A}$ e $\tilde{b}$ tais que $\left|\left(A - \tilde{A}\right)_{i,j}\right| \leq (\Delta_A)_{i,j}$, $\left|b_i - \tilde{b}_i\right| \leq (\Delta_b)_i$ e $\tilde{A}\tilde{x} = \tilde{b}$. Isto equivale a afirmar que existem matrizes δA e δb, $\tilde{A} = A + \delta A$ e $\tilde{b} = b + \delta b$, sujeitas às desigualdades $\left|(\delta A)_{i,j}\right| \leq (\Delta_A)_{i,j}$, $|(\delta b)_i| \leq (\Delta_b)_i$. Então,

$$\begin{aligned}
|(r(\tilde{x}))_i| &= |b_i - (A\tilde{x})_i| \\
&= \left|\tilde{b}_i - (\delta b)_i - \left(\left(\tilde{A} - \delta A\right)\tilde{x}\right)_i\right| \\
&= |-(\delta b)_i + (\delta A\tilde{x})_i| \\
&\leq |(\delta b)_i| + |(\delta A\tilde{x})_i| \\
&\leq (\Delta_b)_i + \sum_j (\Delta_A)_{i,j} |\tilde{x}_j| .
\end{aligned}$$

Vamos agora supor que a desigualdade (2.1.13) é válida. A partir da expressão para o resíduo

$$\begin{aligned}
(r(\tilde{x}))_i &= b_i - \sum_j A_{i,j}\tilde{x}_j \qquad (2.1.14) \\
&= \frac{(\Delta_b)_i + \sum_j (\Delta_A)_{i,j} |\tilde{x}_j|}{(\Delta_b)_i + \sum_j (\Delta_A)_{i,j} |\tilde{x}_j|} \left(b_i - \sum_j A_{i,j}\tilde{x}_j\right) \\
&= \gamma_i (\Delta_b)_i + \sum_j \left(\gamma_i \frac{|\tilde{x}_j|}{\tilde{x}_j} (\Delta_A)_{i,j}\right) \tilde{x}_j \\
&= -(\delta b)_i + \sum_j (\delta A)_{i,j} \tilde{x}_j, \qquad (2.1.15)
\end{aligned}$$

onde $(\delta b)_i = -\gamma_i (\Delta_b)_i$, $(\delta A)_{i,j} = \gamma_i \frac{|\tilde{x}_j|}{\tilde{x}_j} (\Delta_A)_{i,j}$, $\gamma_i = \frac{b_i - \sum_k A_{i,k}\tilde{x}_k}{(\Delta_b)_i + \sum_k (\Delta_A)_{i,k} |\tilde{x}_k|}$ e devido à desigualdade (2.1.13),$|\gamma_i| \leq 1$. Compondo (2.1.14) e (2.1.15) temos

$$\sum_j (A + \delta A)_{i,j} \tilde{x}_j = (b + \delta b)_i ,$$

ou seja

$$\tilde{A}\tilde{x} = \tilde{b},$$

onde

$$\left|\left(\tilde{b}-b\right)_i\right| = |\delta b_i| = |-\gamma_i|\,(\Delta_b)_i \leq (\Delta_b)_i$$

e

$$\left|\left(\tilde{A}-A\right)_{i,j}\right| = \left|(\delta A)_{i,j}\right| = \left|\gamma_i \frac{|\tilde{x}_j|}{\tilde{x}_j}\right| (\Delta_A)_{i,j} \leq (\Delta_A)_{i,j}\,.$$

■

2.2 Refinamento iterativo

O refinamento iterativo de soluções é um método intermediário entre os métodos diretos e os iterativos. Consiste em calcular uma sequência de correções para a solução aproximada de um sistema de equações lineares.

Vamos supor então que $\mathbf{x}^{(0)}$ é uma solução aproximada do sistema

$$\mathrm{A}\mathbf{x} = \mathbf{b}, \tag{2.2.1}$$

calculada através de um método numérico direto como, por exemplo, a eliminação gaussiana. A multiplicação matricial de A pela aproximação $\mathbf{x}^{(0)}$, realizada através de operações aritméticas em ponto flutuante, resulta em um vetor $\mathbf{b}^{(0)}$ diferente da representação em ponto flutuante de $\mathbf{b}$. Para determinar a representação mais próxima à exata no sistema de ponto flutuante utilizado, será necessário obter um vetor ε tal que $\mathbf{x} = \mathbf{x}^{(0)} + \varepsilon$. Ou seja , $\varepsilon = \mathbf{x} - \mathbf{x}^{(0)}$ e assim,

$$\begin{aligned} \mathrm{A}\varepsilon &= \mathrm{A}\mathbf{x} - \mathrm{A}\mathbf{x}^{(0)} \\ &= \mathbf{b} - \mathbf{b}^{(0)} \\ &= \mathbf{r}^{(0)}, \end{aligned}$$

o vetor ε é solução do sistema $\mathrm{A}\varepsilon = \mathbf{r}^{(0)}$, onde $\mathbf{r}^{(0)}$ é o resíduo da solução aproximada $\mathbf{x}^{(0)}$ com relação à solução exata.

A solução numérica do sistema $\mathrm{A}\varepsilon = \mathbf{r}^{(0)}$ estará sujeita às mesmas condições da equação original (2.2.1), portanto podemos esperar obter diretamente um vetor $\varepsilon^{(0)}$ que é solução aproximada para o vetor que corrige a solução $\mathbf{x}^{(0)}$. A correção dessa solução aproximada com o vetor $\varepsilon^{(0)}$ determina a aproximação $\mathbf{x}^{(1)}$ para a solução de (2.2.1) . Esse procedimento é iterado sucessivamente de modo que a $(k+1)$-ésima solução iterada, $\mathbf{x}^{(\mathbf{k+1})}$, é dada pela soma

$$\mathbf{x}^{(\mathbf{k+1})} = \mathbf{x}^{(\mathbf{k})} + \varepsilon^{(k)},$$

onde $\varepsilon^{(k)}$ é o vetor solução do sistema

$$\mathrm{A}\varepsilon^{(k)} = \mathbf{r}^{(\mathbf{k})}$$

e o resíduo $\mathbf{r}^{(\mathbf{k})} := \mathbf{b} - \mathbf{b}^{(\mathbf{k})} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(\mathbf{k})}$, para $k = 0, 1, \ldots$

Quanto á estabilidade numérica desse método, é possível demonstrar que dada uma matriz A, não singular (e portanto, com um condicionamento finito $\left\|\mathrm{A}^{-1}\right\| \|\mathrm{A}\|$), a sequência $\left\{\mathbf{x}^{(\mathbf{i})}\right\}_{\mathbf{i}}$

calculada através de operações em um sistema de ponto flutuante, converge para a representação da solução exata nesse sistema se:

1. a precisão for suficientemente alta

2. os resíduos forem calculados em precisão dupla, antes de serem arredondados para a precisão em que as demais operações são realizadas.

Exemplo 16: Seja o sistema de equações lineares representado pela matriz completa

$$\begin{pmatrix} 0{,}913 & 0{,}659 & 0{,}254 \\ 0{,}781 & 0{,}564 & 0{,}217 \end{pmatrix}.$$

Vamos utilizar o método da eliminação gaussiana para determinar uma aproximação para a solução no sistema de ponto flutuante $F(10, 5, -20, 20)$ com arredondamento por truncamento.

De acordo com o algoritmo vamos realizar a seguinte operação elementar:

$$\begin{aligned} \begin{pmatrix} L_1 \\ (-(0{,}781 \oslash 0{,}913) \otimes L_1) \oplus L_2 \end{pmatrix} &= \begin{pmatrix} L_1 \\ (-0{,}85542 \otimes L_1) \oplus L_2 \end{pmatrix} \\ &= \begin{pmatrix} 0{,}913 & 0{,}659 & 0{,}254 \\ 0 & 2{,}8 \cdot 10^{-4} & -2{,}7 \cdot 10^{-4} \end{pmatrix}, \end{aligned}$$

Obtemos assim a aproximação $\mathbf{x}^{(0)} = \begin{pmatrix} 0{,}97421 & -0{,}96428 \end{pmatrix}^T$. Devemos agora calcular o resíduo em precisão dupla, ou seja, no sistema $F(10, 10, -20, 20)$ e no final arredondamos para o sistema original $F(10, 5, -20, 20)$:

$$\begin{aligned} \mathbf{r}^{(0)} &= \mathbf{b} - \mathrm{A}\mathbf{x}^{(0)} = \begin{pmatrix} 0{,}254 \\ 0{,}217 \end{pmatrix} \ominus \begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}781 & 0{,}564 \end{pmatrix} \otimes \begin{pmatrix} 0{,}97421 \\ -0{,}96428 \end{pmatrix} \\ &= \begin{pmatrix} 0{,}254 \\ 0{,}217 \end{pmatrix} \ominus \begin{pmatrix} 0{,}25399321 \\ -0{,}21700409 \end{pmatrix} \\ &= \begin{pmatrix} 6{,}79 \cdot 10^{-6} \\ -4{,}09 \cdot 10^{-6} \end{pmatrix}. \qquad \text{(nesse caso não houve necessidade de arredondar)} \end{aligned}$$

A primeira correção, $\varepsilon^{(0)}$ é solução do sistema com matriz completa

$$\begin{pmatrix} 0{,}913 & 0{,}659 & 6{,}79 \cdot 10^{-6} \\ 0{,}781 & 0{,}564 & -4{,}09 \cdot 10^{-6} \end{pmatrix}.$$

Antes de realizar o processo de eliminação gaussiana,lembremos que, por definição, a matriz quadrada formada pelas duas primeiras colunas da matriz e a mesma do sistema original. Portanto já conhecemos a sua forma triangular, basta aplicar a operação elementar

$\begin{pmatrix} L_1 \\ (-0{,}85542 \otimes L_1) \oplus L_2 \end{pmatrix}$ à última coluna. Teremos então

$$\begin{pmatrix} 0{,}913 & 0{,}659 & 6{,}79 \cdot 10^{-6} \\ 0 & 2{,}8 \cdot 10^{-4} & -9{,}8983 \cdot 10^{-6} \end{pmatrix}$$

que implica $\varepsilon^{(0)} = \begin{pmatrix} 2{,}5523 \cdot 10^{-2} \\ -3{,}5351 \cdot 10^{-2} \end{pmatrix}$ e $\mathbf{x}^{(1)} = \begin{pmatrix} 0{,}99973 \\ -0{,}99963 \end{pmatrix}$.

O resíduo da nova aproximação é dado por

$$\begin{aligned} \mathbf{r}^{(1)} &= \mathbf{b} - \mathrm{A}\mathbf{x}^{(1)} = \begin{pmatrix} 0.254 \\ 0.217 \end{pmatrix} \ominus \begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}781 & 0{,}564 \end{pmatrix} \otimes \begin{pmatrix} 0{,}99973 \\ -0{,}99963 \end{pmatrix} \\ &= \begin{pmatrix} 2{,}68 \cdot 10^{-6} \\ 2{,}19 \cdot 10^{-6} \end{pmatrix}. \end{aligned}$$

A nova correção $\varepsilon^{(2)}$ é solução do sistema $\mathrm{A}\varepsilon^{(2)} = \mathbf{r}^{(1)}$ representado pela matriz completa

$$\begin{aligned} &\begin{pmatrix} 0{,}913 & 0{,}659 & 2{,}68 \cdot 10^{-6} \\ 0{,}781 & 0{,}564 & 2{,}19 \cdot 10^{-6} \end{pmatrix} \\ \sim\ &\begin{pmatrix} 0{,}913 & 0{,}659 & 2{,}68 \cdot 10^{-6} \\ 0 & 2{,}8 \cdot 10^{-4} & -1{,}025 \cdot 10^{-7} \end{pmatrix} \end{aligned}$$

que implica $\varepsilon^{(2)} = \begin{pmatrix} 2{,}6716 \cdot 10^{-4} \\ -3{,}6607 \cdot 10^{-4} \end{pmatrix}$ e $\mathbf{x}^{(2)} = \begin{pmatrix} 0{,}99999 \\ -0{,}99999 \end{pmatrix}$.

O resíduo associado à nova aproximação ainda não é nulo:

$$\begin{aligned} \mathbf{r}^{(2)} &= \mathbf{b} - \mathrm{A}\mathbf{x}^{(2)} = \begin{pmatrix} 0{,}254 \\ 0{,}217 \end{pmatrix} \ominus \begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}781 & 0{,}564 \end{pmatrix} \otimes \begin{pmatrix} 0{,}99999 \\ -0{,}99999 \end{pmatrix} \\ &= \begin{pmatrix} 2{,}54 \cdot 10^{-6} \\ 2{,}17 \cdot 10^{-6} \end{pmatrix}. \end{aligned}$$

A nova correção $\varepsilon^{(3)}$ é solução do sistema $\mathrm{A}\varepsilon^{(3)} = \mathbf{r}^{(2)}$ representado pela matriz completa

$$\begin{aligned} &\begin{pmatrix} 0{,}913 & 0{,}659 & 2{,}54 \cdot 10^{-6} \\ 0{,}781 & 0{,}564 & 2{,}17 \cdot 10^{-6} \end{pmatrix} \\ \sim\ &\begin{pmatrix} 0{,}913 & 0{,}659 & 2{,}54 \cdot 10^{-6} \\ 0 & 2{,}8 \cdot 10^{-4} & -2{,}7 \cdot 10^{-9} \end{pmatrix} \end{aligned}$$

que implica $\varepsilon^{(2)} = \begin{pmatrix} 9{,}8955 \cdot 10^{-6} \\ -9{,}6428 \cdot 10^{-6} \end{pmatrix}$ e $\mathbf{x}^{(3)} = \begin{pmatrix} 0{,}99999 \\ -0{,}99999 \end{pmatrix}$. Note que $\mathbf{x}^{(3)} = \mathbf{x}^{(2)}$, então a partir desse ponto não conseguimos melhorar a aproximação.

De acordo com o método e o sistema de ponto flutuante utilizado, a solução numérica

do sistema é dada pelo vetor $\begin{pmatrix} 0{,}99999 \\ -0{,}99999 \end{pmatrix}$. Sabemos que a solução exata é $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. A pequena discrepância é consequência da escolha que realizamos para o arredondamento nas operações de ponto flutuante. Note que o número 1 possui duas representações no sistema posicional de base decimal, uma dada pelo algarismo 1 e outra dada por $0{,}99\ldots \equiv 0{,}\bar{9}$. A partir dessa última representação, de acordo com o sistema de ponto flutuante $F(10, 5, -20, 20)$ com arredondamento por truncamento, o número 1 seria armazenado como $0{,}99999 \cdot 10^0$.

2.3 Métodos iterativos

A solução dos sistemas lineares de n equações e n incógnitas através de métodos diretos envolve um número da ordem de n^3 operações em ponto flutuante e, como estudamos, dependendo da natureza das equações, os erros de arredondamento nessas operações podem ser grandes.

Nesta seção vamos estudar métodos em que a solução pode ser aproximada com um número menor de operações, nas quais, os erros de arredondamento estejam sob controle (numericamente estável) .

A estratégia consiste em desenvolver uma regra para construir a sequência de aproximações $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ a partir de uma aproximação inicial $\mathbf{x}^{(0)}$ e garantir a convergência dessa sequência para a solução do sistema.

ponto de partida na consiste em decompor a matriz não singular A no sistema

$$\mathrm{A}\mathbf{x} = \mathbf{b},$$

sob a forma

$$\mathrm{A} = \mathrm{D} - \mathrm{U} - \mathrm{L},$$

onde D é diagonal (e não singular), U e L são respectivamente matrizes triangulares superior e inferior sem a diagonal principal. Dada uma tal escolha, o sistema assume a forma

$$\mathrm{D}\mathbf{x} - (\mathrm{U} + \mathrm{L})\mathbf{x} = \mathbf{b}$$

ou seja,

$$\mathrm{D}\mathbf{x} = \mathbf{b} + (\mathrm{U} + \mathrm{L})\mathbf{x}. \tag{2.3.1}$$

A partir de uma aproximação inicial $\mathbf{x}^{(0)}$ vamos construir a sequência $\left\{\mathbf{x}^{(\mathbf{k})}\right\}_{\mathbf{k}\in\mathbb{N}}$ como solução da seguinte modificação do sistema (2.3.1): no lado direito da equação, utilizamos uma aproximação conhecida para $\mathbf{x}$ e assumimos que a solução do sistema resultante fornece a aproximação seguinte. Ou seja,

$$\mathrm{D}\mathbf{x}^{(\mathbf{k+1})} = \mathbf{b} + (\mathrm{U} + \mathrm{L})\mathbf{x}^{(\mathbf{k})}, \tag{2.3.2}$$

para $\mathbf{k} = 0, 1, \ldots$

Vamos denominar $\varepsilon^{(k)} = \mathbf{x} - \mathbf{x}^{(\mathbf{k})}$, como o erro da k-ésima aproximação com relação à solução exata. Então, a partir de (2.3.2), temos que

$$\begin{aligned}
\mathbf{x^{(k+1)}} &= \mathrm{D}^{-1}\mathbf{b} + \mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\mathbf{x^{(k)}} \\
&= \mathrm{D}^{-1}\left(\mathrm{A}\mathbf{x}\right) - +\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\mathbf{x^{(k)}} \\
&= \mathrm{D}^{-1}\left(\mathrm{D}\mathbf{x} - (\mathrm{U}+\mathrm{L})\mathbf{x}\right) + \mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\mathbf{x^{(k)}} \\
&= \mathbf{x} - \mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\left(\mathbf{x} - \mathbf{x^{(k)}}\right)
\end{aligned}$$

e a partir dessa última igualdade temos

$$\mathbf{x} - \mathbf{x^{(k+1)}} = \mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\left(\mathbf{x} - \mathbf{x^{(k)}}\right) \qquad \Longrightarrow \qquad \varepsilon^{(k+1)} = \mathrm{M}\varepsilon^{(k)}, \tag{2.3.3}$$

onde $\mathrm{M} = \mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})$.

Uma condição necessária e suficiente para que $\lim_{\mathbf{k}\to\infty} \varepsilon^{(k)} = \mathbf{0} \in \mathbb{R}^n$ para qualquer $\varepsilon^{(0)} \in \mathbb{R}^n$ é que M seja tal que $\lim_{\mathbf{k}\to\infty} \mathbf{M^k} = \mathbf{0} \in \mathbb{M}_{n\times n}$. Pois, de acordo com (2.3.3), dada uma aproximação inicial com erro $\varepsilon^{(0)}$, após k iterações, o erro da $(k+1)$-ésima aproximação será

$$\varepsilon^{(k)} = \mathrm{M}\varepsilon^{(k-1)} = \mathrm{M}\mathrm{M}\varepsilon^{(k-2)} = \ldots = \mathrm{M}^k\varepsilon^{(0)}.$$

Matrizes que satisfazem esse limite são denominadas *matrizes convergentes*. Portanto, quando decompomos a matriz não singular A na forma $\mathrm{A} = \mathrm{D} - \mathrm{U} - \mathrm{L}$, devemos realizar um escolha que satisfaça os seguintes critérios:

1. A matriz D deve ser não singular.
2. A matriz $\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})$ deve ser convergente.

O seguinte teorema permite caracterizar uma matriz como convergente ou não, a a partir dos seus autovalores.

Teorema 2.3.1 (**matrizes convergentes**)
Seja $\mathrm{M} \in \mathbb{M}_{n\times n}$ *uma matriz quadrada.* M *é convergente, ou seja,* $\lim_{\mathbf{k}\to\infty} \mathbf{M^k} = \mathbf{0} \in \mathbb{M}_{n\times n}$, *se e somente se o valor absoluto de todos os seus autovalores for estritamente menor do que a unidade.*

O seguinte corolário é muito útil também:

Corolário 2.3.2 (**matrizes convergentes - condição suficiente**)
Uma matriz quadrada M *é convergente se para uma norma matricial qualquer,* $\|\mathrm{M}\| < 1$.

A importância do corolário deve-se ao fato de que existem normas matriciais cujo cálculo é muito simples. De acordo com o corolário, se para qualquer uma dessas normas, a norma de $\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})$ for estritamente menor do que a unidade, então o método será convergente, caso contrário nada podemos afirmar (a não ser que conheçamos os autovalores da matriz). Duas normas que possuem a característica de serem calculadas facilmente são as normas $\|\cdot\|_1$ e $\|\cdot\|_\infty$.

Uma escolha muito natural para a decomposição consistem em eleger B como a matriz diagonal formada pelos elementos na diagonal de A, essa escolha determina o método conhecido como Método de Jacobi.

2.3.1 Método de Jacobi

Dado o sistema de equações lineares

$$\mathrm{A}\mathbf{x} = \mathbf{b},$$

onde A é uma matriz quadrada, $n \times n$, não singular e com elementos $(\mathrm{A})_{i,j} = a_{i,j}$, o método de Jacobi consiste na iteração

$$\mathbf{x}^{(\mathrm{k}+1)} = \mathrm{D}^{-1}\left(\mathbf{b} + (\mathrm{U}+\mathrm{L})\mathbf{x}^{(\mathrm{k})}\right), \qquad \mathbf{k} = 0, 1, \ldots, \tag{2.3.4}$$

onde $\mathbf{x}^{(0)}$ é uma aproximação inicial e B é a matriz diagonal formada pelos elementos da diagonal de A, portanto $(D)_{i,j} = \delta_{i,j}a_{i,j}$ e (U+L) = D−A. Em termos das componentes das aproximações da solução, o método assume a forma da iteração

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j}x_j^{(k)}\right) \tag{2.3.5}$$

para $i = 1, 2, \ldots n$ a cada $k = 0, 1, \ldots$.

Exemplo 17: Seja o seguinte sistema de equações lineares

$$\begin{cases} 5x_1 - x_2 + 3x_3 &= -2 \\ 11x_1 - 21x_3 &= 1 \\ x_1 + 8x_2 + 6x_3 &= 0 \end{cases}.$$

A forma matricial desse sistema, $Ax = b$, implica a seguinte matriz A

$$A = \begin{pmatrix} 5 & -1 & 3 \\ 11 & 0 & -21 \\ 1 & 8 & 6 \end{pmatrix}.$$

O método de Jacobi não pode ser aplicado à equação com a matriz A pois sua diagonal principal é singular: há um zero na diagonal principal, $(A)_{2,2} = 0$. Isto deixa a expressão (2.3.5) indefinida quando $i = 2$.

Esta dificuldade pode ser contornada através da permutação da 2ª e 3ª linha da matriz completa $[A|b]$

$$\left[\tilde{A}|\tilde{b}\right] = \begin{pmatrix} 5 & -1 & 3 & -2 \\ 1 & 8 & 6 & 0 \\ 11 & 0 & -21 & 1 \end{pmatrix}.$$

Note que a troca de linha não altera a solução do sistema mas leva a uma nova matriz $\tilde{A}$ cuja diagonal principal não possui singularidades. O método de Jacobi pode então ser aplicado. A

sequência de aproximações $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ dada pelo método de Jacobi a partir da aproximação inicial $x^{(0)} = \begin{pmatrix} x_1^{(0)} & x_2^{(0)} & x_3^{(0)} \end{pmatrix}^T$ é construída a partir da relação de recorrência

$$\begin{cases} x_1^{(k+1)} &= \frac{1}{5}\left(-2 + x_2^{(k)} - 3x_3^{(k)}\right) \\ x_2^{(k+1)} &= \frac{1}{8}\left(-x_1^{(k)} - x_3^{(k)}\right) \\ x_3^{(k+1)} &= -\frac{1}{21}\left(1 - 11x_1^{(k)}\right) \end{cases}$$

para $k = 0, 1, \ldots$

Aqui cabe uma observação. A troca de linhas da matriz completa não altera a solução de um sistema de equações lineares, porém a convergência do método depende da forma da matriz completa. Ou seja, o comportamento de convergência é potencialmente diferente.

De acordo com o corolário sobre condição suficiente para convergência, o método de Jacobi será convergente se $\left\|\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\right\|_\infty < 1$. Os elementos da matriz dada pelo produto D^{-1}(U+L) são $\left(\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\right)_{i,j} = (1-\delta_{i,j})\dfrac{a_{i,j}}{a_{i,i}}$ (o termo $(1-\delta_{i,j})$ indica que os elementos da diagonal de $\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})$ são nulos). Então podemos concluir através da norma matricial $\|\cdot\|_1$ que se

$$\left\|\mathrm{D}^{-1}(\mathrm{U}+\mathrm{L})\right\|_\infty < 1 \iff \max_{1\le i\le n} \sum_{\substack{j=1 \\ j\neq i}} \left|\frac{a_{i,j}}{a_{i,i}}\right| < 1 \tag{2.3.6}$$

o método converge. A segunda desigualdade em (2.3.6) equivale à condição de "dominância diagonal estrita" da matriz A.

Definição 2.3.3 (matriz diagonal dominante e matriz diagonal dominante estrita). *Seja* $\mathrm{M} \in \mathbb{M}_{n\times n}$ *uma matriz quadrada.* M *é denominada "diagonal dominante" se* $|(M)_{i,i}| \geq \sum_{\substack{j=1 \\ j\neq i}} |(M)_{i,j}|$ *para todo* $i = 1, 2, \ldots n$. *Se a desigualdade* $|(M)_{i,i}| > \sum_{\substack{j=1 \\ j\neq i}} |(M)_{i,j}|, i = 1, 2 \ldots, n$ *for válida então* M *é denominada "diagonal dominante estrita".*

Além da dominância diagonal estrita, outro critério aplicável a matrizes irredutíveis[6][7] garante a convergência do método. Se A for uma matriz irredutível diagonal dominante então a convergência do método de Jacobi também é garantida (a demonstração está no 7º capítulo de Ortega, J.M. "Numerical Analysis: a second course", vol. 3 da coleção "Classics in applied mathematics" editado pela SIAM, Philadelphia 1990).

[6] Uma matriz quadrada M é irredutível se existir uma matriz de permutação P tal que $\mathrm{P}^{-1}\mathrm{MP}$ é uma matriz bloco triangular superior, ou seja, uma matriz quadrada da forma $\begin{pmatrix} \mathrm{B} & \mathrm{E} \\ 0 & \mathrm{D} \end{pmatrix}$ onde B e D são matrizes quadradas. Uma matriz $n \times n$ é irredutível se e somente se para dois índices quaisquer $i, j = 1, 2, \ldots, n$ existir uma sequência de componentes não nulas de M da forma $(\mathrm{M})_{i,i_1}(\mathrm{M})_{i_1,i_2}(\mathrm{M})_{i_2,i_3}\ldots(\mathrm{M})_{i_k,j}$.

[7] Uma matriz de permutação P é uma matriz quadrada que possui uma única componente igual a 1 em cada linha e coluna com todas as demais iguais a 0.

Definição 2.3.4 (matriz diagonal dominante irredutível). *Seja* $\mathrm{M} \in \mathbb{M}_{n\times n}$ *uma matriz quadrada irredutível.* M *é denominada "diagonal dominante irredutível" se for diagonal dominante e além disso* $|(M)_{i,i}| > \sum_{\substack{j=1 \\ j\neq i}} |(M)_{i,j}|$ *para algum* $i = 1, 2, \ldots n$.

2.3.2 Método Gauss–Seidel

O método de Gauss-Seidel consiste em uma pequena modificação do método de Jacobi. Nesse último, para encontrar a aproximação $\mathbf{x^{(k+1)}}$, de acordo com a iteração (2.3.5), devemos conhecer todos as componentes da aproximação anterior $\mathbf{x^{(k)}}$. No método de Gauss-Seidel, levamos em conta que durante a iteração, parte das componentes da próxima aproximação já são conhecidas e são também incorporadas no cálculo das demais. Assim, no método Gauss-Seidel a iteração para cálculo das componentes da $(k+1)$-ésima aproximação é dada por

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{1,1}}\left(b_1 - \sum_{j=2}^{n} a_{1,j}x_j^{(k)}\right), \\
x_i^{(k+1)} &= \frac{1}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}\right), \qquad i = 2, 3, \ldots, n-1, \\
x_n^{(k+1)} &= \frac{1}{a_{n,n}}\left(b_n - \sum_{j=1}^{n-1} a_{n,j}x_j^{(k+1)}\right).
\end{aligned} \tag{2.3.7}$$

Pela estrutura da iteração, podemos notar que, ao contrário do que ocorre no método de Jacobi, não há necessidade de armazenar os dados de todas as componentes da k-ésima aproximação para obter a seguinte.

Exemplo 18: Dado o mesmo sistema do exemplo (17) e matriz a $\left[\tilde{A}|\tilde{b}\right]$:

$$\left[\tilde{A}|\tilde{b}\right] = \begin{pmatrix} 5 & -1 & 3 & -2 \\ 1 & 8 & 6 & 0 \\ 11 & 0 & -21 & 1 \end{pmatrix}.$$

A sequência de aproximações $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ dada pelo método Gauss-Seidel a partir da aproximação inicial $x^{(0)} = \begin{pmatrix} x_1^{(0)} & x_2^{(0)} & x_3^{(0)} \end{pmatrix}^T$ é construída a partir da relação de recorrência

$$\begin{cases}
x_1^{(k+1)} &= \frac{1}{5}\left(-2 + x_2^{(k)} - 3x_3^{(k)}\right) \\
x_2^{(k+1)} &= \frac{1}{8}\left(-x_1^{(k+1)} - x_3^{(k)}\right) \\
x_3^{(k+1)} &= -\frac{1}{21}\left(1 - 11x_1^{(k+1)}\right)
\end{cases}$$

para $k = 0, 1, \ldots$

Com relação à convergência desse método, podemos notar pela forma das iteradas em (2.3.7)

que a condição será distinta daquela obtida no método de Jacobi. No entanto, para o seu desenvolvimento é necessário expressar $\mathbf{x}^{(k+1)}$ em função de $\mathbf{x}^{(k)}$. A partir da segunda e terceira linhas de (2.3.7) multiplicadas respectivamente por $a_{i,i}$ e $a_{n,n}$temos

$$\begin{aligned} a_{i,i}x_i^{(k+1)} &= b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}, \qquad i = 2, 3, \ldots, n-1, \\ a_{n,n}x_n^{(k+1)} &= b_n - \sum_{j=1}^{n-1} a_{n,j}x_j^{(k+1)}. \end{aligned}$$

Agora passando as componentes $x_j^{(k+1)}$ para o lado esquerdo

$$\begin{aligned} \sum_{j=1}^{i} a_{i,j}x_j^{(k+1)} &= b_i - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}, \qquad i = 2, 3, \ldots, n-1, \\ \sum_{j=1}^{n} a_{n,j}x_j^{(k+1)} &= b_n. \end{aligned} \tag{2.3.8}$$

A primeira linha de (2.3.7) e as duas de (2.3.8) em notação matricial assumem a forma

$$(\mathrm{D} - \mathrm{L})\mathbf{x}^{(k+1)} = \mathbf{b} + \mathrm{U}\mathbf{x}^{(k)}$$

e portanto

$$\mathbf{x}^{(k+1)} = (\mathrm{D} - \mathrm{L})^{-1}\left(\mathbf{b} + \mathrm{U}\mathbf{x}^{(k)}\right), k = 0, 1, \ldots \tag{2.3.9}$$

Se a matriz A for diagonal dominante estrita ou diagonal dominante irredutível então a sequência dada pelo método Gauss-Seidel converge para a solução do sistema independente da aproximação inicial. A demonstração dessa condição suficiente é análoga àquela do método de Jacobi.

2.4 Exemplos comentados

Sistema massa–mola em equilíbrio estático

Desejamos obter a solução de um sistema formado por três corpos: m_1, m_2 e m_3. Sujeitos à ação de uma força elástica de acordo com o diagrama.

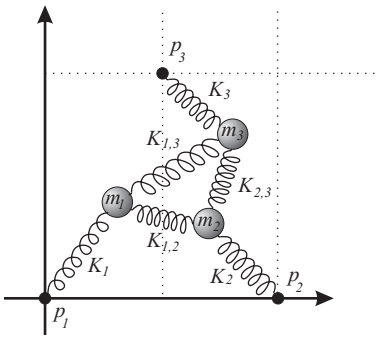


Figura 2.4.1: Sistema massa-mola. As constantes de mola valem $K_1 = K_2 = K_3 = 2Nm^{-1}$ e $K_{1,2} = K_{2,3} = K_{1,3} = 1Nm^{-1}$. As coordenadas das extremidades fixas são dadas pelos vetores $p_1 = \begin{pmatrix} 0 & 0 \end{pmatrix}^T$, $p_2 = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$ e $p_3 = \begin{pmatrix} 2^{-1} & 1 \end{pmatrix}^T$ em unidades S.I.

Para minimizar a quantidade de termos nas expressões, consideraremos que todas as quantidades estão em unidades S.I. e não faremos referências às mesmas.

Nosso objetivo é obter os valores das componentes do vetor posição dos corpos m_1, m_2 e m_3. Simbolizaremos os vetores posição desses corpos, respectivamente por $r_1, r_2, r_3 \in \mathbb{R}^2$. Assim, as incógnitas do problema são as componentes desses três vetores :$r_{1,x}$, $r_{1,y}$, $r_{2,x}$, $r_{2,y}$, $r_{3,x}$ e $r_{3,y}$. São seis incógnitas, portanto devem haver seis equações que as relacione.

As equações advém da exigência de que o sistema esteja em equilíbrio estático. Nesse caso, a resultante das forças sobre cada corpo é nula. As únicas forças em jogo são as forças elásticas cuja forma é dada pela lei de Hooke.

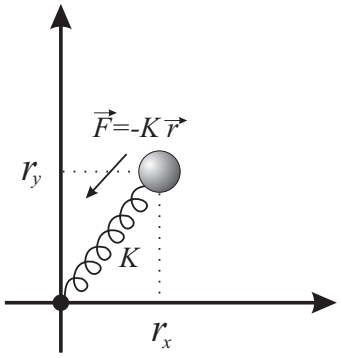


Figura 2.4.2: Lei de Hooke.

A resultante das forças sobre o corpo m_1 (equação vetorial)

$$-K_1 (r_1 - p_1) - K_{1,2} (r_1 - r_2) - K_{1,3} (r_1 - r_3) = \mathbf{0} \in \mathbb{R}^2,$$

onde $p_1 = \begin{pmatrix} 0 & 0 \end{pmatrix}^T \in \mathbb{R}^2$ é a coordenada da extremidade fixa da mola K_1. Como a equação acima é uma equação que envolve vetores, ela deve ser analisada componente a componente.

Componente x:

$$-K_1 r_{1,x} + K_1 p_{1,x} - K_{1,2} (r_{1,x} - r_{2,x}) - K_{1,3} (r_{1,x} - r_{3,x}) = 0,$$

que equivale a[8]

$$(K_1 + K_{1,2} + K_{1,3})\, r_{1,x} - K_{1,2} r_{2,x} - K_{1,3} r_{3,x} = K_1 p_{1,x}.$$

Substituindo o valor das constantes de mola e da componente x da extremidade fixa,

$$4r_{1,x} - r_{2,x} - r_{3,x} = 0. \tag{2.4.1}$$

Componente y:

$$-K_1 r_{1,y} + K_1 p_{1,y} - K_{1,2}\,(r_{1,y} - r_{2,y}) - K_{1,3}\,(r_{1,y} - r_{3,y}) = 0,$$

que equivale a

$$(K_1 + K_{1,2} + K_{1,3})\, r_{1,y} - K_{1,2} r_{2,y} - K_{1,3} r_{3,y} = K_1 p_{1,y}.$$

Substituindo o valor das constantes de mola e da componente x da extremidade fixa,

$$4r_{1,y} - r_{2,y} - r_{3,y} = 0. \tag{2.4.2}$$

Colecionando as equações (2.4.1), (2.4.2) e as demais equações para as componentes x e y dos corpos m_2 e m_3, obtém-se o seguinte sistema

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

A estrutura matricial fica mais claramente evidenciada se renomearmos as variáveis na forma

$$x_1 := r_{1,x},\ x_2 := r_{1,y},\ x_3 := r_{2,x},\ x_4 := r_{2,y},\ x_5 := r_{3,x},\ x_6 := r_{3,y}.$$

[8] A equação vetorial para esse sistema é facilmente generalizada para o caso de um sistema com um número arbitrário de massas conectadas por molas, parte delas com extremidades fixas em uma região do $\mathbb{R}^n$.

Seja $N \subset \mathbb{N}$, um conjunto de índices para as massas. Cada massa está ligada a pelo menos uma das outras massas e, naturalmente, não está ligada a ela mesma por uma mola. Assim, a cada $i \in N$, existe um conjunto não vazio de índices $I_i \subset N$ que guarda os índices das outras massas ligadas à massa i. Do mesmo modo, cada massa i pode estar ligada por uma mola a uma extremidade fixa. Assim, a cada $i \in N$, existe um conjunto $\Phi_i \subset \mathbb{N}$ que guarda o índice de uma extremidade fixa. Vamos supor ainda, que cada massa pode estar sob a ação de uma força externa (por exemplo, a gravidade).

A equação para a posição de equilíbrio $r_i \in \mathbb{R}^n$ da massa $i \in N$ é dada por

$$\left(\sum_{l\in\Phi_i} K_{i,l} + \sum_{j\in I_i} K_{i,j}\right) r_i - \sum_{j\in I_i} K_{i,j} r_j = \sum_{l\in\Phi_i} K_{i,l} p_l + F_i,$$

onde $K_{i,j} = K_{j,i}$ é a constante de mola que liga a massa i à massa j ou extremidade fixa j, p_l é o vetor posição da extremidade fixa l e F_i á força externa sobre a massa i.

Nesse caso, o sistema assume a forma

$$Ax = b,$$

onde

$$A = \begin{pmatrix} 4 & 0 & -1 & 0 & -1 & 0 \\ 0 & 4 & 0 & -1 & 0 & -1 \\ -1 & 0 & 4 & 0 & -1 & 0 \\ 0 & -1 & 0 & 4 & 0 & -1 \\ -1 & 0 & -1 & 0 & 4 & 0 \\ 0 & -1 & 0 & -1 & 0 & 4 \end{pmatrix} \quad \text{e} \quad b = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix}. \tag{2.4.3}$$

Verificamos imediatamente que a matriz A é estritamente diagonal dominante (linhas e colunas). Portanto, além dos métodos diretos, podemos utilizar tanto o método de Jacobi quanto Gauss-Seidel com garantia de convergência.

Dada a sua estrutura, criamos no Scilab a matriz de coeficientes e coluna das constantes do sistema (2.4.3) com a sequência de comandos

```
A=4*eye(6,6);
for i=1:4
  A(i,i+2)=-1;
  A(i+2,i)=-1;
end;
for i=1:2
  A(i,i+4)=-1;
  A(i+4,i)=-1;
end;
b=[0 0 2 0 1 2]';
```

A representação matricial do sistema depende da ordem com que são atribuídas as componentes dos vetores posição. Por exemplo, a escolha alternativa

$$\tilde{x}_1 := r_{1,x},\ \tilde{x}_2 := r_{2,x},\ \tilde{x}_3 := r_{3,x},\ \tilde{x}_4 := r_{1,y},\ \tilde{x}_5 := r_{2,y},\ \tilde{x}_6 := r_{3,y},$$

leva ao sistema

$$\begin{pmatrix} 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -1 & -1 \\ -1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 \\ -1 & -1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 \end{pmatrix} \cdot \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \\ \tilde{x}_5 \\ \tilde{x}_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix}.$$

A permutação da segunda e terceira linhas, seguida pela permutação da terceira e quinta linhas da matriz completa associada produz um sistema com representação matricial

$$\tilde{A}\tilde{x} = \tilde{b}$$

onde

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 & 0 \\ -1 & -1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -1 & -1 \\ 0 & 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & -1 & 4 \end{pmatrix} \quad \text{e} \quad b = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}. \tag{2.4.4}$$

A novidade aqui é que esse novo sistema pode ser resolvido como dois sistemas independentes.

O sistema representado por (2.4.4) possui uma matriz de coeficientes que é bloco diagonal. Isto permite separar o sistema original em dois sistemas independentes. Porém, dada a estrutura de A nesse ordenamento, a matriz de coeficientes desses dois sistemas é igual, portanto, para obter a solução numérica no Scilab, construiremos a matriz de coeficientes e a matriz de constantes a partir dos comandos

```
A=-
ones(3,3)+5*eye(3,3);
b=[0 2 1;0 0 2]';
```

A solução numérica é obtida a partir de algum dos seguintes comandos:

- `A\b`
- `gseidel([A b],tol=0)`
- `gpp([A b])`
- `jacobi_solv([A b],tol=0)`

A solução exata é dada por

$$r_1 = \begin{pmatrix} 0{,}3 \\ 0{,}2 \end{pmatrix}, \qquad r_2 = \begin{pmatrix} 0{,}7 \\ 0{,}2 \end{pmatrix} \quad \text{e} \quad r_3 = \begin{pmatrix} 0{,}5 \\ 0{,}6 \end{pmatrix}$$

Rede de resistores

Considere o circuito elétrico descrito pela figura seguinte

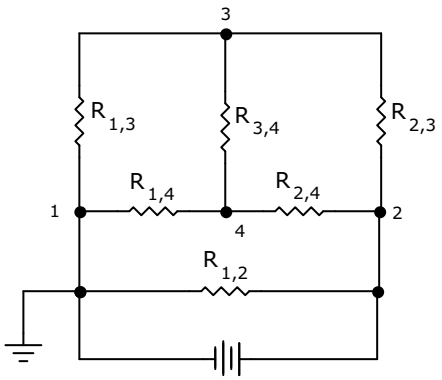


Figura 2.4.3: Circuito elétrico. A fonte é de $1V$ e as resistências $R_{2,4} = R_{3,4} = 1\Omega$, $R_{2,3} = R_{1,4} = 5\Omega$ e $R_{1,2} = R_{1,3} = 10\Omega$.

Uma vez estabelecidos os valores das resistências e da voltagem na fonte, as correntes e tensões em cada terminal são determinados por um sistema de equações lineares que é resultado da aplicação da Lei de Ohm e da Lei de Kirchoff para correntes.

A notação adotada para tensões e correntes é seguinte, V_i simboliza a tensão no nó i e $I_{i,j}$, a corrente no sentido do nó i para o nó j que atravessa um resistor cujas extremidades estão ligadas a esses nós. Note que, de acordo com a notação, $I_{i,j} = -I_{j,i}$.

A partir dessa notação, a Lei de Ohm assume a forma

$$V_i - V_j = RI_{i,j}, \tag{2.4.5}$$

onde R é a resistência do resistor.

A Lei de Kirchoff assume a forma

$$\sum_{i \in J} I_{i,j} = 0, \tag{2.4.6}$$

onde J é conjunto formado pelos índices dos nós ligados ao nó j.

De acordo com a estrutura do circuito, convencionalmente $V_1 = 0$ pois o nó 1 está ligado ao terra, V_2 é determinado pela fonte e $I_{1,2} = -\dfrac{V_2}{R_{1,2}}$. As demais tensões e correntes são incógnitas do problema. Não é difícil verificar que são sete incógnitas: V_3, V_4, $I_{1,4}$, $I_{2,4}$, $I_{3,4}$, $I_{1,3}$ e $I_{2,3}$. A partir de (2.4.5) e (2.4.6), as sete relações entre as incógnitas são dadas pelas equações

$$\left.\begin{aligned} V_1 - V_4 &= R_{1,4}I_{1,4} \\ V_2 - V_4 &= R_{2,4}I_{2,4} \\ V_3 - V_4 &= R_{3,4}I_{3,4} \\ V_1 - V_3 &= R_{1,3}I_{1,3} \\ V_2 - V_3 &= R_{2,3}I_{2,3} \end{aligned}\right\} \text{Lei de Ohm.} \qquad \left.\begin{aligned} I_{1,4} + I_{2,4} + I_{3,4} &= 0 \\ I_{1,3} + I_{2,3} - I_{3,4} &= 0 \end{aligned}\right\} \begin{array}{l}\text{Lei de Kirchoff} \\ \text{para os nós 4 e 3.}\end{array}$$

A descrição desse sistema na forma matricial

$$Ax = b$$

depende de um ordenamento das incógnitas como componentes de x. A escolha

$$x_1 := V_3,\ x_2 := V_4,\ x_3 := I_{1,3},\ x_4 := I_{1,4},\ x_5 := I_{2,3},\ x_6 := I_{2,4},\ x_7 := I_{3,4}, \tag{2.4.7}$$

resulta nas seguintes matrizes

$$A = \begin{pmatrix} 0 & -1 & 0 & -R_{1,4} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & -R_{2,4} & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -R_{3,4} \\ -1 & 0 & -R_{1,3} & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -R_{2,3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & -1 \end{pmatrix} \quad \text{e} \quad b = \begin{pmatrix} 0 \\ -V_2 \\ 0 \\ 0 \\ -V_2 \\ 0 \\ 0 \end{pmatrix}.$$

A escolha alternativa

$$x_1 := I_{1,4},\ x_2 := I_{2,4},\ x_3 := I_{2,3},\ x_4 := I_{1,3},\ x_5 := I_{3,4},\ x_6 := V_3,\ x_7 := V_4, \tag{2.4.8}$$

resulta nas seguintes matrizes

$$A = \begin{pmatrix} R_{1,4} & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & R_{2,4} & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & R_{2,3} & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & R_{1,3} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & R_{3,4} & -1 & 1 \end{pmatrix} \quad \text{e} \quad b = \begin{pmatrix} 0 \\ V_2 \\ V_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Esta última escolha produz uma matriz de coeficientes sem elementos nulos na diagonal principal, mas isto não é suficiente para produzir uma matriz diagonal dominante.

A solução numérica com seis dígitos para o sistema no ordenamento (2.4.7) é dada por

$$x \approx \begin{pmatrix} 7{,}74194D-01 \\ 8{,}06452D-01 \\ -7{,}74194D-02 \\ -1{,}61290D-01 \\ 4{,}51613D-02 \\ 1{,}93548D-01 \\ -3{,}22581D-02 \end{pmatrix}.$$

A partir da solução sabemos que a tensão no nó 3 vale $\approx 0{,}774194V$ e no nó 4 vale $\approx 0{,}806452V$. Além disso, a corrente que sai da fonte é dada por $I_{2,1} + I_{2,3} + I_{2,4} = 0.1 + x_5 + x_6 \approx 0{,}338710A$. Portanto a rede de resistores possui um resistência equivalente de $\approx 2{,}95238\Omega$.

2.5 Exercícios

1) Utilize o método de Gauss com pivotamento parcial para encontrar a solução do seguinte sistema de equações lineares

$$\begin{cases} 4x + 4y &= 20{,}5 \\ 7x + 6{,}99y &= 34{,}97 \end{cases}$$

no sistema $F(10, 5, -10, 10)$. Verifique que o sistema possui solução exata que pode ser representada nesse sistema de ponto flutuante e realize as operações de refinamento.

2) Utilize o método de Gauss com pivotamento parcial para encontrar a solução do seguinte sistema de equações lineares

$$\begin{cases} 4x + 4y &= 20 \\ 7x + 6{,}9y &= 34{,}7 \end{cases}$$

no sistema $F(10, 5, -10, 10)$. Verifique que o sistema possui solução exata que pode ser representada nesse sistema de ponto flutuante e realize as operações de refinamento.

3) Seja o sistema $A\,\mathbf{x} = \mathbf{b}$, onde

$$A = \begin{pmatrix} 0{,}5 & 0{,}4 \\ 0{,}3 & 0{,}25 \end{pmatrix}.$$

Considere que conheçamos a aproximação $\tilde{\mathbf{b}} = \begin{pmatrix} 0{,}2 & 1 \end{pmatrix}^T$ para o lado direito da equação. Se os coeficientes da aproximação $\tilde{\mathbf{b}}$ são exatos até o terceiro dígito após a vírgula encontre uma estimativa para o erro relativo na resposta $\tilde{\mathbf{x}}$, isto é, para $\dfrac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_1}{\|\mathbf{x}\|_1}$.

4) Seja o sistema com matriz completa (o mesmo utilizado no exemplo da seção refinamento iterativo)

$$\begin{pmatrix} 0{,}913 & 0{,}659 & 0{,}254 \\ 0{,}781 & 0{,}564 & 0{,}217 \end{pmatrix}.$$

Encontre a solução através do método de eliminação gaussiana com variáveis no sistema de ponto flutuante $F(10, 5, -20, 20)$ e arredondamento par. Utilize a refinação iterativa até que a aproximação não mais seja modificada.

5) Utilize o método de Jacobi e o método de Gauss-Seidel para esse mesmo sistema. O que você pode dizer sobre a convergência dos métodos? (trabalhe com a maior precisão possível).

6) Seja o sistema de equações algébricas lineares cuja matriz completa é dada por A:

$$A = \begin{pmatrix} 10 & -3 & \alpha & 4 & \beta \\ 0 & 9 & \beta & 5 & \alpha \\ 1 & -1 & 10 & 2 & \alpha \\ 0 & 1 & 2 & -\beta & 1 \end{pmatrix}.$$

Determine o intervalo de valores que α e β podem assumir de modo que o sistema possa ser resolvido através do método Gauss-Seidel com garantia de convergência.

7) Seja A uma matriz simétrica $n \times n$ de componentes $[A]_{i,j}$:

$$[A]_{i,j} = \begin{cases} \alpha & , \ i = j \\ \\ r^{|i-j|} & , \ i \neq j \end{cases} .$$

Para quais valores de α o sistema

$$A\mathbf{x} = \mathbf{b}$$

pode ser resolvido via Jacobi com garantias de convergência?

8) Um determinado problema possui solução na forma de um sistema de equações lineares com n ($n \geq 11$) variáveis x_i, $i = 1, 2, \ldots, n$ e n equações

$$\begin{array}{lcrcl}
\text{Primeira equação} & : & 8\,x_1 + 4\,x_3 + 2\,x_5 - x_{n-2} - x_n & = & 1 \\
i = 2, \ldots, 5 & : & x_{i-1} - 4\,x_i + x_{i+1} - x_{n-i} + x_n & = & \cos\left(3\dfrac{i-1}{n-1}\pi\right) \\
i = 6, \ldots, n-5 & : & x_{i-3} - 2\,x_{i-2} + 8\,x_i - 2\,x_{i+2} + x_{i+3} & = & \cos\left(3\dfrac{i-1}{n-1}\pi\right) \\
i = n-4, \ldots, n-1 & : & x_1 - x_2 + x_{i-1} - 4x_i + x_{i+1} & = & \cos\left(3\dfrac{i-1}{n-1}\pi\right) \\
\text{Última equação} & : & -x_1 - x_3 + 2\,x_{n-4} + 4\,x_{n-2} + 8\,x_n & = & -1
\end{array}$$

Considere o caso $n = 50$ e justifique se há garantia de convergência pelo método de Jacobi ou Gauss-Seidel se a matriz for criada a partir do ordenamento indicado acima. Determine o valor de $(x_1 - x_n)^2/2$ com seis dígitos de precisão.

9) Um determinado problema possui solução na forma de um sistema de equações lineares com n variáveis x_i, $i = 1, 2, \ldots, n$ e n equações

$$\begin{array}{lcrcl}
\text{Primeira equação} & : & n\,x_1 + (n-1)\,x_2 + \ldots + 2\,x_{n-1} + x_n & = & b_1 \\
i = 2, \ldots, n-1 & : & 2x_{i-1} - 4\,x_i + 3x_{i+1} & = & b_i \\
\text{Última equação} & : & x_1 + 2\,x_2 + \ldots + (n-1)\,x_{n-1} + n\,x_n & = & b_n
\end{array}$$

Considere o caso $n = 50$ e determine uma estimativa na norma 2 para o erro relativo na solução se o erro relativo dos coeficientes b_i for igual a 10^{-9}.

10) O problema de contorno

$$\begin{cases} y'' + x^2 y = \exp(-x), & x \in (0, 2) \\ y(0) = 0, \quad y(2) = 0 \end{cases}$$

quando discretizado nos n pontos $x_i = (i-1)h$, $i = 1, 2, \ldots n$ com $h = \dfrac{2}{n-1}$, dá origem ao sistema de n equações lineares

$$\begin{cases}
y_1 = 0 \\
y_{i+1} + \left(h^4(i-1)^2 - 2\right) y_i + y_{i-1} = h^2 \exp\left(-(i-1)h\right), \qquad i = 2, 3, \ldots, n-1 \\
y_n = 0
\end{cases}$$

Se o número de pontos n for igual a $2^{-p} + 1$ para algum inteiro positivo $p \leq 13$, os termos da matriz de coeficientes serão representados exatamente nos registros de ponto flutuante de 64 bits.

A partir do condicionamento da matriz, determine uma aproximação com quatro dígitos para o erro relativo da solução em norma 2 considerando que o erro relativo nos termos do lado direito são da ordem de 10^{-16} e $n = 257$.

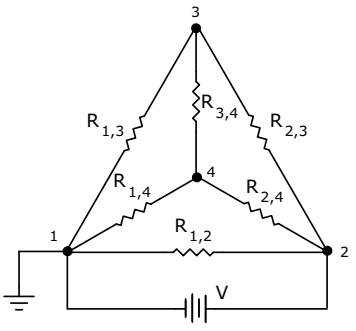
11) Um problema de contorno discretizado em n pontos uniformente ao longo do domínio $(0, 1)$ dá origem ao seguinte sistema de n equações lineares

$$\begin{cases} y_1 &= 0 \\ y_{i+1} + \left(h^2 + h^2\cos\left(2(i-1)h\right) - 2\right) y_i + y_{i-1} &= h^2 \exp\left(-(i-1)h\right), i = 2, 3, \ldots, n-1 \\ y_n &= 0 \end{cases}$$

onde $h = \dfrac{1}{n-1}$. A presença do termo $h^2\cos\left(2(i-1)h\right)$ impede que uma grande quantidade de termos não nulos da matriz dos coeficientes do sistema sejam representados exatamente por registros de ponto flutuante. Em situações semelhantes, a relação entre os erros relativos δx, δA, δb e o condicionamento $\kappa_\alpha(A)$ (em norma α) no sistema de equações $Ax = b$ satisfaz a desigualdade $\delta x \leq \dfrac{\kappa_\alpha(A)}{1 - \kappa_\alpha(A)\delta A}\left(\delta A + \delta b\right)$ se $\kappa_\alpha(A)\delta A < 1$. Trabalhando com a norma ∞, determine uma estimativa superior com quatro dígitos para o erro relativo δx no caso em que $n = 150$, $\delta b \lessapprox 10^{-16}$ e $\delta A \lessapprox 10^{-16}$.

12) Considere o circuito elétrico descrito pelo diagrama abaixo. As resistências são dadas por $R_{2,4} = R_{3,4} = 15\Omega$, $R_{2,3} = R_{1,4} = 10\Omega$ e $R_{1,2} = R_{1,3} = 30\Omega$ e a voltagem $V = 3$V. A notação adotada para tensões e correntes é seguinte, V_i simboliza a tensão no nó i e $I_{i,j}$, a corrente no sentido do nó i para o nó j que atravessa um resistor cujas extremidades estão ligadas a esses nós. Note que, de acordo com a notação, $I_{i,j} = -I_{j,i}$. De acordo com as leis de Ohm e Kirchoff para circuitos, a corrente $I_{2,1} = V/R_{1,2}$ e as demais tensões e correntes satisfazem as equações ao lado. Determine o valor absoluto da maior corrente (6 dígitos na resposta).

$$\begin{cases} V_4 &= -R_{1,4}I_{1,4} \\ 3 - V_4 &= R_{2,4}I_{2,4} \\ V_3 - V_4 &= R_{3,4}I_{3,4} \\ V_3 &= -R_{1,3}I_{1,3} \\ 3 - V_3 &= R_{2,3}I_{2,3} \\ I_{1,4} + I_{2,4} + I_{3,4} &= 0 \\ I_{1,3} + I_{2,3} - I_{3,4} &= 0 \end{cases}$$



13) Considere o seguinte conjunto de reações químicas em um recipiente fechado e a volume constante:

Reação 1	$2S_1 + 3S_3$	$\rightarrow$	$S_3 + 2S_5$
Reação 2	$S_2 + S_3$	$\rightarrow$	$3S_4$
Reação 3	$S_1 + 3S_5$	$\rightarrow$	$2S_4 + 2S_6$
Reação 4	$S_4 + S_6$	$\rightarrow$	$3S_7$

A matriz estequiométrica N desse sistema é construída de modo que $(N)_{i,j}$ representa o coeficiente da susbtância j na i-ésima reação (o sinal é positivo se a substância estiver do lado direito e negativo caso contrário, se uma mesma substância estiver dos dois lado, faz-se a soma dos coeficientes levando em conta o sinal), por exemplo, a 1ª linha é da forma [-2,0,-2,0,2,0,0] e a 2ª linha [0,-1,-1,3,0,0,0]. A partir da taxa de produção e/ou consumo de cada uma das substâncias é possível determinar a taxa de reação de cada uma das reações a partir da solução do sistema de equações lineares $\left(N * N^T\right) * \tau = N * R$, onde τ é a matriz coluna com a taxa de cada uma das equações e R é a matriz coluna com as taxas de produção/consumo das substâncias. Determine qual das reações possui a menor taxa (e o seu valor com três dígitos), dado que R =[-1.5;-0.91;-1.2;1.6;-0.24;0.88;1.3] mol/(l s).

3 Equações não lineares

Vamos estudar métodos numéricos para resolver o seguinte problema. Dada uma função f contínua, real e de uma variável, queremos encontrar uma solução x^* que satisfaça a equação não linear:

$$f(x^*) = 0. \tag{3.0.1}$$

Em geral a equação (3.0.1) não pode ser resolvida exatamente, isto é, a solução x^* não pode ser descrita a partir de uma combinação finita de operações algébricas simples $(+, -, /, \times, \exp, \log)$ e funções elementares (polinômios, razão entre polinômios, potências racionais, e as funções transcendentais: exp,log, trigonométricas, hiperbólicas). Há casos em que a própria função f não é conhecida explicitamente: pode ser definida a partir de uma série infinita, ou a partir de uma integral ou ainda ser solução de uma equação diferencial. nesses casos utilizamos métodos numéricos para resolver a equação.

Idealmente, poderíamos dividir o procedimento nas seguintes etapas: inicialmente devemos encontrar uma região de interesse onde possam existir soluções da equação; em seguida, quando possível, isolar os intervalos que contém apenas 1 solução; feito isso, determinamos pelo menos 1 aproximação inicial $x^{(0)}$ da solução (de acordo com o método utilizado, pode ser necessário utilizar mais de uma aproximação inicial) para cada intervalo; finalmente, a partir das aproximações iniciais, o método numérico consiste na construção de uma sequência $\{x^{(0)}\}_{n=0}^{\infty}$ que converge para a solução , isto é,

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

solução da equação (3.0.1). Portanto os métodos numéricos para encontrar a solução de equações não lineares são *métodos iterativos*. A cada iteração, utilizamos um subconjunto das aproximações $x^{(n-1)}, x^{(n-2)}, \ldots, x^{(0)}$, obtidas anteriormente, para determinar a próxima aproximação $x^{(n)}$.

Taxa de convergência e critérios de parada

A velocidade com que uma sequência de aproximações converge para a solução é conhecida como "taxa de convergência". Se uma sequência de aproximações $\left\{x^{(n)}\right\}_{n=0}^{\infty}$ converge para a solução x^*, diz-se que a sequência "converge linearmente" para x^*se existir uma constante real $L \in (0, 1)$, denominada "taxa de convergência", tal que

$$\lim_{n \to \infty} \frac{\left|x^{(n+1)} - x*\right|}{\left|x^{(n)} - x*\right|} = L.$$

Tecnicamente, o limite não fornece informação sobre o comportamento de subsequências finitas da sequência de aproximações mas permite estimar e comparar a velocidade com que sequências

construídas por diferentes métodos converge para a solução.

As extremidades do intervalo de valores que a taxa de convergência pode assumir estão associados a situações especiais:

- Se a razão $\dfrac{\left|x^{(n+1)} - x*\right|}{\left|x^{(n)} - x*\right|}$ não é constante e o limite corresponde a $L = 1$, então diz-se que a sequência "converge sublinearmente". Se, além disso,

$$\lim_{n\to\infty} \frac{\left|x^{(n+1)} - x^{(n)}\right|}{\left|x^{(n)} - x^{(n-1)}\right|} = 1,$$

  então diz-se que a sequência "converge logaritmicamente".

- Se o limite corresponde a $L = 0$, então diz-se que a sequência "converge superlinearmente". Neste caso, é possível fazer uma distinção entre as formas de convergência. Se existir um real $q > 1$ tal que

$$\lim_{n\to\infty} \frac{\left|x^{(n+1)} - x*\right|}{\left|x^{(n)} - x*\right|^q} > 0,$$

  diz-se que a sequência "converge com ordem q". Quando $q = 2$, diz-se que a convergência é "quadrática", quando $q = 3$, diz-se que é "cúbica" e assim por diante.

Os métodos iterativos, em geral, fornecem uma maneira de se obter a solução exata do problema mediante a repetição infinita de operações realizadas sobre os números reais (ou complexos) o que é operacionalmente impraticável. Na construção concreta das aproximações através de operações em máquina será necessário determinar o momento de interromper a sequência quando as aproximações satisfizerem algum critério. Nos casos em que a convergência é linear ou superlinear é possível estabelecer um critéria de parada da forma

$$\left|x^{(n+1)} - x^{(n)}\right| \leq tol * \left|x^{(n+1)}\right|, \tag{3.0.2}$$

onde *tol* é um fator positivo menor que a unidade denominado "tolerância".

Uma outra alternativa para critério de parada consiste na desigualdade

$$|x_{n+1} - x_n| \geq |x_n - x_{n-1}| \tag{3.0.3}$$

para n maior ou igual a um número mínimo de iteradas. A justificativa para o seu uso é o fato de que devido aos erros de arredondamento nas operações de ponto flutuante, a sequência de aproximações pode entrar em um ciclo onde os mesmos valores são obtidos indefinidamente. Caso isso ocorra, a desigualdade (3.0.3) será satisfeita.

Se o parâmetro *tol* for suficientemente pequeno, a partir de um determinado índice na sequência de aproximações, $x^{(n+1)}$ pode estar tão próximo de $x^{(n)}$ a ponto de serem vizinhos no sistema de ponto flutuante ao qual pertencem. Considere um sistema de ponto flutuante com com n dígitos binários para o significando. Seja $x = (1{,}d_1 d_2 \ldots d_{n-1})_2 \times 2^E$ um elemento desse sistema, os seus vizinhos mais próximos distam $2^{-(n-1)} \times 2^E$, ou $2^{-(n-2)} \times 2^E$ quando x é da forma

$(1{,}00\ldots0)_2 \times 2^{E+1}$. A partir da multiplicação da desigualdade

$$1 \leq (1{,}d_1 d_2 \ldots d_{n-1})_2$$

pelas distâncias entre os vizinhos mais próximos de x temos

$$\begin{aligned} 2^{-(n-1)} \times 2^E \leq \quad & 2^{-(n-1)}\,(1{,}d_1 d_2 \ldots d_{n-1})_2 \times 2^E \quad = 2^{-(n-1)}x \\ & \text{ou} \\ 2^{-(n-2)} \times 2^E \leq \quad & 2 \times 2^{-(n-1)}\,(1{,}d_1 d_2 \ldots d_{n-1})_2 \times 2^E \quad = 2 \times 2^{-(n-1)}x. \end{aligned}$$

O que permite concluir que a distância entre um ponto flutuante não nulo e os seus vizinhos mais próximos é menor ou igual a

$$2 \times 2^{-(n-1)}|x|.$$

Os pontos flutuantes utilizados por vários programas numéricos correspondem aos "doubles" definido pelo padrão IEEE-754, pontos flutuantes com 64 bits dos quais 52 são utilizados para representar 53 dígitos binários, ou seja, $n = 53$. Neste caso, o termo 2^{-52}, denominado pelo Scilab (e também pelo padrão IEEE-754) como "epsilon de máquina", simbolizado por ε_m, quantifica o limite superior para o erro relativo nas operações aritméticas com pontos flutuantes. A partir dessa descrição, a distância entre dois pontos flutuantes vizinhos é limitada superiormente por

$$2\varepsilon_m|x|.$$

Dessa forma, se o parâmetro tol em (3.0.2) for menor que $2\varepsilon_m$, a parada na construção da sequência de aproximações ocorreria próxima da capacidade da máquina em diferenciar dois pontos flutuantes, ou seja, o resultado seria equivalente a escolher $tol = 0$.

Condicionamento

O comportamento de f próximo ao zero x^* e as limitações impostas pela aritmética de máquina dão origem aos problemas de condicionamento na determinação de aproximações para x^*. Na situação em que $f'(x^*)$ possui valores muito pequenos, uma pequena incerteza no valor de f induz grandes incertezas no valor de x. A questão é quantificar esse efeito, o que pode ser alcançado através da análise da série de Taylor para a função f.

Vamos supor que f é contínua em um intervalo fechado com extremidades x e x^* e diferenciável no intervalo aberto com as mesmas extremidades, então, de acordo com o teorema de Taylor[1] existe um ξ no intervalo aberto tal que

$$f(x) = f(x^*) + (x - x^*)\,f'(\xi).$$

Antes de continuar a análise, é util considerar a definição de multiplicidade dos zeros.

[1] O teorema admite diversas formas para o termo correspondente ao erro em uma série de Taylor. Uma referência que trata desse assunto é

- Apostol, T. M. "Calculus : One-Variable Calculus with an Introduction to Linear Algebra, Vol. 1, Ed. 2", (1967), seções 7.6 e 7.7, página 283.

Definição 3.0.1 (Multiplicidade dos zeros). *Seja f uma função p vezes continuamente diferenciável em alguma vizinhança de x^*, raiz da equação $f(x) = 0$. Dizemos que x^* possui multiplicidade p se*

$$0 < \lim_{x \to x^*} \frac{|f(x)|}{|x - x^*|^p} < \infty.$$

É usual denominar os zeros de multiplicidade 1 como "zeros simples".

Se x^* é um zero simples de f, existe uma constante positiva L , tal que $|f'| \geq L$ no intervalo aberto e além disso, por definição $f(x^*) = 0$, então

$$x - x^* = \frac{f(x)}{f'(\xi)} \implies |x - x^*| \leq \frac{|f(x)|}{L}. \tag{3.0.4}$$

Em máquina, f é calculada numericamente através de operações em ponto flutuante. O resultado é uma outra função, $\tilde{f}$, que difere de f pela presença de um termo aditivo δ que contém os erros cometidos nas operações necessárias para avaliar o valor de f

$$\tilde{f}(x) = f(x) + \zeta(x), \tag{3.0.5}$$

onde[2] a função ζ é limitada superiormente em valor absoluto por uma constante δ, ou seja, $|\zeta| \leq \delta$. Se $x^{(n)}$ é uma aproximação para a solução x^*e seu valor é o de um registro de um sistema de ponto flutuante com a propriedade

$$\tilde{f}\left(x^{(n)}\right) = 0, \tag{3.0.6}$$

então de acordo com a aritmética de máquina, $x^{(n)}$ é uma solução da equação $f(x) = 0$. Combinando (3.0.4) , (3.0.5) e (3.0.6) chega-se à desigualdade

$$\left|x^{(n)} - x^*\right| \leq \frac{\delta}{L}.$$

Assim, a razão $\dfrac{\delta}{L}$ fornece informação sobre a acurácia com que é possível obter a solução da equação não linear. Se a derivada em torno de um zero simples assume valores muito pequenos, a tarefa de determinar uma aproximação com boa acurácia pode ser muito prejudicada.

No caso de um zero de multiplicidade p há um resultado semelhante. Dado um inteiro $p > 1$ e uma função f de classe $\mathcal{C}^{p-1}$ no intervalo fechado com extremidades x e x^*, cuja derivada de ordem p está definida no intervalo aberto de mesmas extremidades, existe um ξ contido nesse intervalo aberto tal que

$$f(x) = \frac{1}{p!} f^{(p)}(\xi) \left(x - x^*\right)^p,$$

já que, por hipótese, $f(x^*) = f'(x^*) = \ldots = f^{(p-1)}(x^*) = 0$. Dessa expressão, segue por argumentos similares que para uma aproximação dada por um registro de ponto flutuante igual

[2] A função $\tilde{f}$ assume valores em um sistema de pontos flutuantes. Em geral, determinar o valor de δ para qualquer função f não é uma tarefa trivial pois esse problema se relaciona ao problema conhecido como "Table Maker's Dilemma". Referência:

- Muller, J.-M.;*et all*. "Handbook of Floating-Point Arithmetic". Springer-Verlag. New York. 2009.

a $x^{(n)}$, se $\left|f^{(p)}\right| > L_p$ no intervalo aberto com extremidades em $x^{(n)}$ e x^*, então

$$\left|x^{(n)} - x^*\right| \leq \left(\frac{\delta}{L_p} p!\right)^{1/p}.$$

Os métodos podem ser separados em três classes principais:

- **Métodos de quebra**: o ponto de partida é encontrar um intervalo que contenha pelo menos 1 solução. Segundo o teorema de Bolzano, basta determinar um intervalo em que a função f muda de sinal. Os métodos de quebra consistem na descrição de como subdividir o intervalo inicial em intervalo cada vez menores que ainda contenham a mesma solução. Nesse caso, a sequência $x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ é formada pelos extremos dos intervalos. A solução numérica será encontrada quando a largura do intervalo em uma m-ésima iteração for pequeno o suficiente para satisfazer as exigências de exatidão.

- **Métodos de ponto fixo**: A sequência $\{x^{(i)}\}_{i=0}^{i=n}$ é construída a partir da sucessiva iteração $x^{(n+1)} = \phi\left(x^{(n)}\right)$. A convergência do método é garantida pelo teorema do ponto fixo, daí o nome dos métodos.

- **Métodos de múltiplos passos**: Uma generalização do método anterior onde a função ϕ depende de mais de uma aproximação anterior, i. e., $x^{(n+1)} = \phi(x^{(n)}, x^{(n-1)}, \ldots, x^{(n-p)})$ para algum $p \geq n$.

3.1 Métodos de quebra

Os métodos de quebra utilizam como primeira aproximação um intervalo que contenha pelo menos 1 solução da equação não linear. As iterações consistem em seguidas subdivisões dos intervalos de maneira que o novo intervalo sempre contenha a solução. O uso do teorema é comum a todos os métodos de quebra, ele fornece condições para que os intervalos contenham pelo menos uma solução para a equação. Apresentamos o teorema sem sua prova.

Teorema 3.1.1 (Bolzano)
Seja $I = [a, b] \in \mathbb{R}$ e uma função $f : I \to \mathbb{R}$ contínua. Então o conjunto imagem $f(I)$ é também um intervalo e $[f(a), f(b)] \subseteq f(I)$ ou $[f(b), f(a)] \subseteq f(I)$.

Portanto, se encontrarmos um intervalo $[a, b]$ tal que, por exemplo, $f(a) < 0$ e $f(b) > 0$, então pelo teorema de Bolzano existe, um ponto $x^* \in [a, b]$ tal que $f(x^*) = 0$.

O que difere os métodos de quebra entre si é a maneira com que os intervalos são subdivididos.

3.1.1 Método da bissecção

O passo inicial no método exige o conhecimento prévio de um intervalo $\left[x^{(0)}, x^{(1)}\right]$ tal que o produto $f\left(x^{(0)}\right) f\left(x^{(1)}\right)$ seja negativo. De acordo com o teorema de Bolzano, há pelo menos uma solução nesse intervalo. A sequência de aproximações $\left\{x^{(0)}, x^{(1)}, x^{(2)}, \ldots\right\}$ é construída de acordo com os seguintes passos:

- As duas aproximações iniciais, $x^{(0)}$ e $x^{(1)}$, são os extremos do intervalo inicial, a partir deles escolhemos o ponto intermediário $x_m = \dfrac{x^{(0)} + x^{(1)}}{2}$ como a nova aproximação.
- A partir de x_m, dividimos o intervalo ao meio, ou seja, o intervalo original é dividido em duas subintervalos de mesmo comprimento: $\left[x^{(0)}, x_m\right]$ e $\left[x_m, x^{(1)}\right]$.
- De acordo com a escolha do intervalo inicial, o sinal do valor numérico de $f\left(x^{(0)}\right)$ é diferente do sinal de $f\left(x^{(1)}\right)$, portanto ao calcular o valor de $f\left(x^{(2)}\right)$, há três possibilidades: $f(x_m) = 0$ e a solução é x_m; o sinal de $f(x_m)$ é igual ao sinal de $f\left(x^{(0)}\right)$ ou o sinal de $f(x_m)$ é igual ao sinal de $f\left(x^{(1)}\right)$. Caso $f(x_m) \neq 0$, entre $\left[x^{(0)}, x_m\right]$ e $\left[x_m, x^{(1)}\right]$, o intervalo que garantidamente possui pelo menos uma solução é aquele cujo produto de f nos extremos for menor que zero. Os novos extremos que satisfaze essa condição são renomeados $x^{(2)}$ e $x^{(3)}$, ou seja,

$$x^{(2)} = \begin{cases} x^{(0)}, & \text{se } f\left(x^{(0)}\right) f(x_m) < 0 \\ x_m, & \text{caso contrário} \end{cases}$$

e

$$x^{(3)} = \begin{cases} x_m, & \text{se } f\left(x^{(0)}\right) f(x_m) < 0 \\ x^{(1)}, & \text{caso contrário.} \end{cases}$$

- O subintervalo resultante possui metade do comprimento do intervalo original, $\left[x^{(0)}, x^{(1)}\right]$ e seus extremos são indicados pelos pontos $x^{(2)}$ e $x^{(3)}$. Esse novo subintervalo é submetido a nova divisão e o procedimento é repetido e os índices dos pontos obtidos é atualizado.

Dessa forma, a partir das aproximações iniciais $x^{(0)}$ e $x^{(1)}$ construiremos uma sequência de intervalos $\{\left[x^{(0)}, x^{(1)}\right], \left[x^{(2)}, x^{(3)}\right], \left[x^{(4)}, x^{(5)}\right], \ldots, \left[x^{(2k)}, x^{(2k+1)}\right], \ldots\}$ cujos comprimentos decrescem com uma razão $1/2$ e os pontos $x^{(2k)}$ e $x^{(2k+1)}$ são determinados a partir das regras

$$x_m = \frac{x^{(2k-2)} + x^{(2k-1)}}{2},$$

$$x^{(2k)} = \begin{cases} x^{(2k-2)}, & \text{se } f\left(x^{(2k-2)}\right) f(x_m) < 0, \\ x_m, & \text{caso contrário,} \end{cases}$$

e

$$x^{(2k+1)} = \begin{cases} x_m, & \text{se } f\left(x^{(2k-2)}\right) f(x_m) < 0, \\ x^{(2k-1)}, & \text{caso contrário,} \end{cases}$$

para $k = 1, 2, \ldots$

Observação 3.1.2. *Se a aproximação inicial $\left[x^{(0)}, x^{(1)}\right]$ for tal que $f\left(x^{(0)}\right) f\left(x^{(1)}\right) > 0$ isto não quer dizer que não exista solução nesse intervalo, apenas o teorema não permite uma conclusão sobre a existência ou não de solução nesse intervalo. Nesse caso é necessário escolher outro intervalo ou então realizar um divisão adicional. Por exemplo, se $f(x) = x(1-x)$, a equação não linear $f(x^*) = 0$ possui soluções $x^* = 0$ e $x^* = 1$, porém $f(-1)f(3) > 0$.*

Devemos adotar um critério de parada no processo de subdivisão dos intervalos. É comum utilizar dois parâmetros: um de valor pequeno, ε, a partir do qual o processo é interrompido se

a desigualdade $\left|x^{(n)} - x^{(n-1)}\right| < \varepsilon$ for satisfeita e um parâmetro inteiro, N, que representa o número máximo de iterações aceitáveis.

Como exemplo do método, vamos estudar a equação não linear para $f(x) = x - e^{-x}$. A solução é dada em termos da função especial W de Lambert[3]:

$$x^* = W(1) = 0{,}56714329040978387\ldots$$

A tabela seguinte ilustra o comportamento dos extremos do intervalo para a equação $x - e^{-x} = 0$ com intervalo inicial $(0{,}0\,;\,1{,}0)$:

iteração i	$x^{(2i)}$	$x^{(2i+1)}$
1	0,5	1,0
2	0,5	0,75
3	0,5	0,625
4	0,5625	0,625
5	0,5625	0,59375
6	0,5625	0,578125

Tabela 3.1: Tabela das primeiras iterações para o método da bissecção.

Após 20 iterações chegamos ao intervalo $[0{,}567142\ldots,\ 0{,}567143\ldots]$. O valor 0,567143 é satisfatório como solução com 6 dígitos exatos.

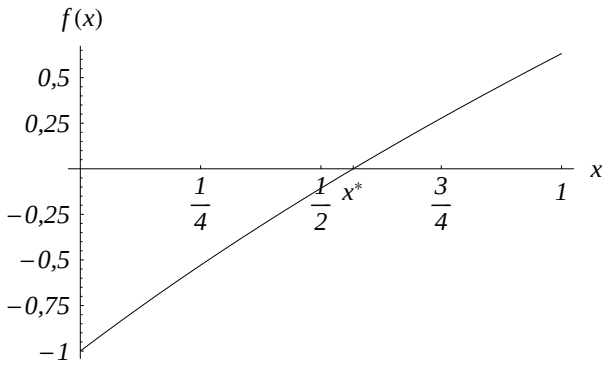
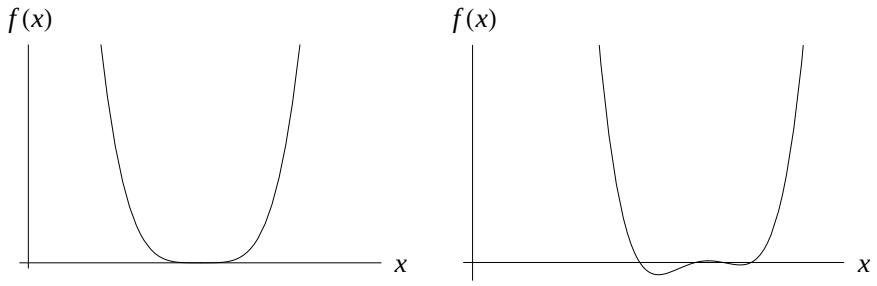


Figura 3.1.1: Gráfico da função $f(x) = x - e^{-x}$, no intervalo $x \in [0, 1]$.

Limitações do método

Se na solução x^*, a primeira derivada não nula da função f for de ordem par então existe uma vizinhança de x^* na qual f possui o mesmo sinal. Veja os gráficos abaixo:

[3]Dado um real $y > -\frac{1}{e}$, a função $W(y)$ é definida como o número real W que satisfaz a equação $y = We^W$. Ou seja, W é a inversa da função xe^x.



A função f na figura à esquerda possui uma região extensa de valores do argumento x em que $f(x)$ é quase nula e além disso, a menos de sua raiz x^*, f é sempre positiva[4]. Nesse caso, apesar de existir uma solução x^*, não há intervalo em que f troca de sinal.

Já a função f na figura à direita possui quatro raízes distintas, porém em um intervalo extenso de valores do argumento x, $f(x)$ é quase nula. Nesse caso, dependendo da precisão utilizada, o método pode identificar apenas algumas das soluções.

3.1.2 Método da falsa posição ou *regula falsi*

A diferença básica entre este método e o método da bisseção está na forma de dividir o intervalo. O método da falsa posição utiliza como ponto intermediário para divisão do intervalo $\left(x^{(0)}, x^{(1)}\right)$, o ponto dado pela intersecção entre o eixo x e a reta que une os pontos $\left(x^{(0)}, f\left(x^{(0)}\right)\right)$ e $\left(x^{(1)}, f\left(x^{(1)}\right)\right)$. A reta que une esses dois pontos possui equação $\rho(x)$:

$$\rho(x) = \frac{f\left(x^{(0)}\right) - f\left(x^{(1)}\right)}{x^{(0)} - x^{(1)}} x + \frac{x^{(0)} f\left(x^{(1)}\right) - x^{(1)} f\left(x^{(0)}\right)}{x^{(0)} - x^{(1)}}.$$

Portanto, o ponto intermediário x_m é dado por

$$x_m = x^{(0)} - \frac{\left(x^{(0)} - x^{(1)}\right)}{f\left(x^{(0)}\right) - f\left(x^{(1)}\right)} f\left(x^{(0)}\right).$$

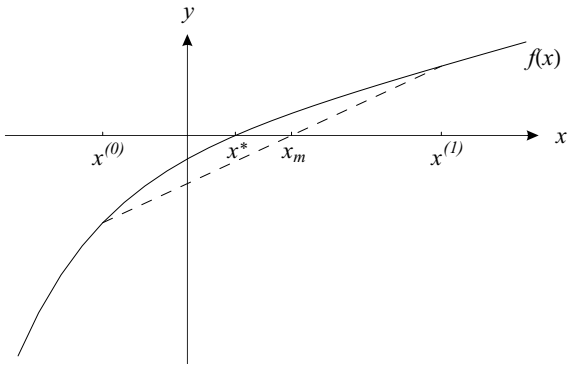


Figura 3.1.2: A reta que une os pontos $\left(x^{(0)}, f\left(x^{(0)}\right)\right)$ e $\left(x^{(1)}, f\left(x^{(1)}\right)\right)$ está pontilhada. Ela cruza o eixo x no ponto que divide o intervalo, x_m

[4]Corresponde ao caso de raízes de multiplicidade par nos polinômios, ou de maneira mais geral, quando existe um natural n tal que $\frac{d^{2j-1}}{dx^{2j-1}} f(x^*) = 0$ para todo $0 \leq j \leq n$ e $\frac{d^{2n}}{dx^{2n}} f(x^*) \neq 0$.

A tabela seguinte ilustra o comportamento dos extremos do intervalo para a equação $x - e^{-x} = 0$ com intervalo inicial $(0.0\,,\ 1.0)$:

iteração i	x_{2i}	x_{2i+1}
1	0,0	0,612699 ...
2	0,0	0,572181 ...
3	0,0	0,567703 ...
4	0,0	0,567206 ...
5	0,0	0,567150 ...
6	0,0	0,567144 ...

Tabela 3.2: Tabela das primeiras iterações para o método da falsa posição.

Após 7 iterações chegamos ao resultado nas mesmas condições (6 dígitos de exatidão) utilizadas no método anterior.

3.2 Métodos de ponto fixo

Os métodos de ponto fixo caracterizam-se por reescrever a equação não linear

$$f(x^*) = 0 \tag{3.2.1}$$

na forma

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

e utilizar o teorema do ponto fixo – que veremos logo adiante – para garantir a convergência da sequência $x^{(n+1)} = \phi\left(x^{(n)}\right)$ para o ponto fixo x^* que é solução de (3.2.1).

Vamos considerar um intervalo $[a, b]$ que contém uma única solução $x^* \in [a, b]$ da equação (3.2.1). Nesse intervalo, definimos a função $\phi : [a, b] \to \mathbb{R}$

$$\phi(x) = x + \gamma(x) f(x),$$

onde $\gamma(x) \neq 0$ no intervalo $[a, b]$. Como γ não se anula em todo intervalo, a equação $\phi(x) = x$ possui x^* como única solução.

A solução x^* será então determinada através da convergência da sequência $\left\{x^{(n)}\right\}_{n=0}^{\infty}$, $\lim_{n\to\infty} x^{(n)} = x^*$. A garantia da convergência é estabelecida pelo teorema do ponto fixo:

Teorema 3.2.1 (ponto fixo)
Seja ϕ uma função contínua em um intervalo $I = [a, b]$ e diferenciável no intervalo aberto (a, b). Se as seguintes condições forem satisfeitas:

- $\phi(I) \subseteq I$, *obs: (a notação indica $\forall x \in I, g(x) \in I$)*
- $\forall x \in I, |\phi'(x)| \leq L < 1$ *obs:(ou seja, ϕ é uma contração)*

Então dado qualquer $x^{(0)} \in I$, existe um único ponto $x^* \in I$ tal que a sequência $x^{(n+1)} = \phi\left(x^{(n)}\right)$ converge para $x^* = \phi\left(x^*\right)$.

Demonstração: A demonstração está estruturada da seguinte forma: inicialmente vamos relacionar os erros absolutos na $(n+1)$-ésima aproximação aos erros absolutos na aproximação anterior e garantir que a cada iteração o erro seja menor. Em seguida trataremos da unicidade de solução em I, para tanto vamos supor que existam duas e concluir por absurdo que isso não é possível.

Trataremos agora da questão da convergência (existência de ponto fixo). Seja a distância entre a $(n+1)$-ésima aproximação , $x^{(n+1)}$, e a solução exata, x^* :

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

A igualdade é válida pois, por definição, $x_{n+1} = \phi(x_n)$ e $x^* = \phi(x^*)$.

Segundo o teorema do valor médio, existe um $c \in \left(x^{(n)}, x^*\right)$ tal que $\left|\phi\left(x^{(n)}\right) - \phi\left(x^*\right)\right| = \left|\phi'(c)\right| \left|x^{(n)} - x^*\right|$. De acordo com as hipóteses, ϕ é tal que $\phi(I) \subseteq I$, então se a aproximação inicial, $x^{(0)}$, pertence a I, $\phi\left(x^{(n)}\right)$ também pertence a esse intervalo. Como $c \in \left(x^{(n)}, x^*\right) \subset I$, segundo as hipóteses temos que $|\phi'(c)| \leq L < 1$, e assim:

$$\begin{aligned} \left|x^{(n+1)} - x^*\right| &= \left|\phi\left(x^{(n)}\right) - \phi\left(x^*\right)\right| \\ &= \left|\phi'(c)\right| \left|x^{(n)} - x^*\right| \\ &\leq L \left|x^{(n)} - x^*\right|. \end{aligned} \tag{3.2.2}$$

Utilizando recursivamente a desigualdade (3.2.2) verificamos que

$$\left|x^{(n+1)} - x^*\right| \leq L \left|x^{(n)} - x^*\right| \leq L^2 \left|x^{(n-1)} - x^*\right| \leq \ldots \leq L^{n+1} \left|x^{(0)} - x^*\right|,$$

portanto

$$\lim_{n\to\infty} L \left|x^{(n+1)} - x^*\right| \leq \lim_{n\to\infty} L^{n+1} \left|x^{(0)} - x^*\right| = \left|x^{(0)} - x^*\right| \lim_{n\to\infty} L^{n+1}.$$

Novamente segundo as hipóteses, $L < 1$ e $\left|x^{(0)} - x^*\right|$ é um número finito pois $x^{(0)}$ e x^* pertencem ao intervalo finito $[a, b]$. Assim $\lim_{n\to\infty} L^{n+1} = 0$ e

$$\lim_{n\to\infty} \left|x^{(n+1)} - x^*\right| = 0.$$

Ou seja, a sequência converge para um $x^* = \phi(x^*)$. Dessa forma, existe pelo menos um ponto x^* no intervalo $[a, b]$ que satisfaz a equação $x^* = \phi(x^*)$. A seguir, vamos verificar que esse ponto é único.

Sejam x^{*1}e x^{*2} dois pontos distintos no intervalo $I = [a, b]$ que satisfazem a equação $x = \phi(x)$, ou seja, $x^{*1} = \phi\left(x^{*1}\right)$ e $x^{*2} = \phi\left(x^{*2}\right)$. Então, de acordo com o teorema do valor médio, existe um $c \in \left(x^{*1}, x^{*2}\right)$ tal que

$$\left|x^{*1} - x^{*2}\right| = \left|\phi'(c)\right| \left|x^{*1} - x^{*2}\right|.$$

Segundo as hipóteses, $x^{*1}, x^{*2} \in I$, dessa forma $\phi'(c) \leq L < 1$, ou seja,

$$\left|x^{*1} - x^{*2}\right| < \left|x^{*1} - x^{*2}\right|,$$

o que é uma contradição.

Portanto, no intervalo $I = [a, b]$ há um e somente um ponto $x^* = \phi(x)$. ■

Observação 3.2.2. *Note que na demonstração do teorema do ponto fixo, é fundamental que a derivada de ϕ* ***seja estritamente menor do que*** *1 em alguma vizinhança I que contém a solução. Caso contrário, se $|\phi'(x)| \leq 1$ em um intervalo I,* ***não podemos excluir a possibilidade*** *de que as iteradas transitem por uma sequência cíclica de pontos sem convergir para a solução x^*, ou mesmo a possibilidade de haver mais de uma solução nesse intervalo. Naturalmente isto não quer dizer que esses comportamentos ocorram sempre que as hipóteses do teorema não forem válidas.*

3.2.1 Método da iteração linear

Trata-se de encontrar uma função ϕ que satisfaça as hipóteses do teorema do ponto fixo para alguma vizinhança em torno da solução x^* da equação $f(x^*) = 0$.

Como a função ϕ é construída a partir de uma outra função $\gamma(x) \neq 0$ em um intervalo que contem a solução de $f(x^*) = 0$, encontrá-la significa determinar $\gamma(x) \neq 0$. A condição de convergência é garantida então pelo teorema do ponto fixo se as suas hipóteses forem satisfeitas.

Vamos considerar o exemplo que já estudamos anteriormente, $f(x) = x - e^{-x}$. Nesse caso $f(x) = 0 \Rightarrow x = e^{-x} = \phi(x)$. Portanto, como por definição, $\phi(x) = x + \gamma(x)f(x)$, no nosso exemplo $\gamma(x) \equiv -1$. Assim $\gamma(x) \neq 0$ para qualquer valor de x. Como $|\phi'(x)| = e^{-x}$, as hipóteses do teorema do ponto fixo são válidas no intervalo[5] $I = (0, +\infty)$, onde $\phi(I) \subset I$ e $|\phi'(x)| < 1$.

Vamos então, escolher como aproximação inicial $x_0 = 0{,}5$. A sequência é dada em seus primeiros termos por

iteração n	x_n
1	0,606531...
2	0,545239...
3	0,579703...
4	0,560065...
5	0,571172...
6	0,565863...

Tabela 3.3: Tabela das primeiras iterações para o método da iteração linear com $\phi(x) = e^{-x}$.

A solução com 6 dígitos exatos é alcançada após 22 iterações.

Uma outra possibilidade para a função ϕ seria a escolha $\phi(x) = -\ln x$ que corresponde a $\gamma(x) = -\dfrac{x + \ln x}{x - e^{-x}}$ que é sempre negativa no intervalo $(0, +\infty)$. No entanto, $|\phi'(x)| = \dfrac{1}{|x|}$ é maior do que a unidade no intervalo $(0, 1)$ que contém a solução e assim, o teorema do ponto fixo não dá garantias de convergência. Podemos perceber que logo nas primeiras iterações, a

[5] Na prática, raramente procuramos garantir a hipótese $\phi(I) \subseteq I$ pois muitas vezes, determinar esse intervalo exatamente equivale a resolver a equação não linear.

sequência toma valores negativos e, dessa forma, como $\phi(x) = -\ln x$, a sequência não estará definida apenas nos números reais. Em particular essa sequência não converge para nenhuma solução de $f(x)$ no plano complexo (a equação possui infinitas soluções lá).

3.2.2 Método Newton–Raphson

A partir da demonstração do teorema do ponto fixo, podemos notar que quanto menor for o limite superior $L < 1$ para o valor absoluto da derivada de ϕ na vizinhança da solução x^* mais rapidamente a sequência converge para a solução da equação não linear. O método de Newton-Raphson é um método iterativo que utiliza essa propriedade da convergência das sequências para garantir uma convergência rápida para a solução a partir do instante que x_{n+1} se aproxima de uma vizinhança suficientemente próxima de x^*. Portanto, a ideia é determinar uma função $\gamma(x)$ tal que ϕ e sua derivada sejam contínuas em algum domínio contínuo que contenha a solução e além disso, $\phi'(x^*) = 0$. Essas hipóteses garantem que, em uma vizinhança próxima de x^*, a função ϕ é tal que $|\phi'| \ll 1$.

Tomando a derivada de ϕ, por definição temos:

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

e em $x = x^*$, solução da equação $f(x^*) = 0$, temos

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

Portanto, a escolha

$$\gamma(x) = -\frac{1}{f'(x)} \tag{3.2.3}$$

implica $\phi'(x^*) = 0$ de maneira que na vizinhança de x^*, $|\phi'|$ assume pequenos valores. A partir da escolha (3.2.3) para a função γ, a função ϕ é dada por

$$\phi(x) = x - \frac{f(x)}{f'(x)}. \tag{3.2.4}$$

Uma outra forma de se obter a fórmula de Newton é realizar uma expansão em série de Taylor em torno de uma aproximação $x^{(n)}$ do zero de f. Para tanto, é necessário que f pertença à classe de funções contínuas em uma vizinhança do zero x^* que contenha também a aproximação $x^{(n)}$ e que a derivada f' exista nessa mesma vizinhança. Por simplicidade, vamos supor que $x^{(n)} > x^*$. Então de acordo com o teorema de Taylor, existe um $\xi \in \left(x^{(n)}, x^*\right)$ tal que a expansão em série de Taylor em torno de $x^{(n)}$ e calculada em x^* é dada por

$$f(x^*) = f\left(x^{(n)}\right) + f'(\xi)\left(x^* - x^{(n)}\right)$$

como $f(x^*) = 0$, os termos podem ser reagrupados na forma

$$x^* = x^{(n)} - \frac{f\left(x^{(n)}\right)}{f'(\xi)}. \tag{3.2.5}$$

Naturalmente, o valor de ξ não é conhecido, porém se $x^{(n)}$ for suficientemente próximo de x^*, o lado direito da expressão (3.2.5), com ξ substituído por $x^{(n)}$, pode ser utilizado como uma nova aproximação $x^{(n+1)}$ para x^*:

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

Esta forma de desenvolver o método de Newton-Raphson permite também analisar a questão do convergência. Mas antes de tratar dessa questão, vamos novamente determinar uma aproximação para o zero real de $f(x) = x - e^{-x}$. Neste caso, a iteração é dada pela função ϕ:

$$\phi(x) = x - \frac{x - e^{-x}}{1 + e^{-x}} = \frac{(x+1)}{1 + e^{x}}.$$

Partindo da aproximação inicial $x_0 = 0{,}5$:

iteração n	x_n
1	0,566311 ...
2	0,567143 ...

Tabela 3.4: Tabela das primeiras iterações para o método Newton-Raphson com $\phi(x) = \frac{(x+1)}{1+e^x}$.

a sequência converge para a solução exata até a 6ª casa decimal em duas iterações. Se utilizarmos $x^{(0)} = 1{,}0$ como aproximação inicial obteríamos o mesmo resultado após três iterações.

Vamos analisar com um pouco mais de detalhe a questão da convergência. Seja $x^{(n+1)}$ um ponto dado pela relação de recorrência (3.2.4) que está próximo da solução x^*, então de acordo com a definição, $|x_{n+1} - x^*| = |\phi(x_n) - \phi(x^*)|$. Estudamos na subseção anterior que se ϕ for contínua e possuir derivada contínua no intervalo aberto entre os pontos x_n e x^*, o teorema do valor médio garante que existe um ponto c nesse intervalo tal que

$$\begin{aligned} \left|x^{(n+1)} - x^*\right| &= \left|\phi\left(x^{(n)}\right) - \phi\left(x^*\right)\right| \\ &= \left|\phi'(c)\right| \left|x^{(n)} - x^*\right|. \end{aligned} \tag{3.2.6}$$

Se x_{n+1} estiver suficientemente próximo de x^* e $f'(x^*) \neq 0$, $\phi'(c)$ será um número pequeno. Dito de uma forma mais exata,

$$\lim_{n\to\infty} \frac{\left|x^{(n+1)} - x^*\right|}{\left|x^{(n)} - x^*\right|} = \left|\phi'\left(x^*\right)\right| = \frac{f\left(x^*\right) f''\left(x^*\right)}{\left(f'\left(x^*\right)\right)^2} = 0, \qquad \text{se } f'\left(x^*\right) \neq 0. \tag{3.2.7}$$

Ou seja, nessa situação, a convergência é mais rápida que a linear.

É possível analisar mais detalhadamente o comportamento da sequência através do Teorema de Taylor, em particular, com a forma de Lagrange para o erro. De acordo com o teorema, se f é uma função com derivada contínua em um aberto que contenha os pontos x^* e $x^{(n)}$ (sem perda de generalidade, vamos supor que $x^{(n)} > x^*$) e f'' existe nesse mesmo intervalo, então existe um

$\xi \in \left(x^*, x^{(n)}\right)$ tal que

$$f\left(x^{(n)}\right) = f\left(x^*\right) + \left(x^{(n)} - x^*\right) f'\left(x^*\right) + \frac{1}{2}\left(x^{(n)} - x^*\right)^2 f''\left(\xi\right).$$

O ponto ξ depende de f, x^* e $x^{(n)}$. Sob as mesmas condições, existe um $\eta \in \left(x^*, x^{(n)}\right)$ tal que

$$f'\left(x^{(n)}\right) = f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right).$$

A partir da relação de recorrência do método e substituindo $f\left(x^{(n)}\right)$ e $f'\left(x^{(n)}\right)$ pelas formas dadas pelo teorema de Taylor,

$$\begin{aligned}
x^{(n+1)} - x^* &= \phi\left(x^{(n)}\right) - \phi\left(x^*\right) \\
&= x^{(n)} - \frac{f\left(x^{(n)}\right)}{f'\left(x^{(n)}\right)} - x^* \\
&= x^{(n)} - x^* - \frac{f\left(x^*\right) + \left(x^{(n)} - x^*\right) f'\left(x^*\right) + \frac{1}{2}\left(x^{(n)} - x^*\right)^2 f''\left(\xi\right)}{f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right)} \\
&= x^{(n)} - x^* - \frac{\left(x^{(n)} - x^*\right) f'\left(x^*\right) + \frac{1}{2}\left(x^{(n)} - x^*\right)^2 f''\left(\xi\right)}{f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right)} \\
&= \left(x^{(n)} - x^*\right)\left(1 - \frac{f'\left(x^*\right) + \frac{1}{2}\left(x^{(n)} - x^*\right) f''\left(\xi\right)}{f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right)}\right) \\
&= \left(x^{(n)} - x^*\right)\left(\frac{\left(x^{(n)} - x^*\right)\left(f''\left(\eta\right) - \frac{1}{2} f''\left(\xi\right)\right)}{f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right)}\right) \\
&= \left(x^{(n)} - x^*\right)^2 \left(\frac{f''\left(\eta\right) - \frac{1}{2} f''\left(\xi\right)}{f'\left(x^*\right) + \left(x^{(n)} - x^*\right) f''\left(\eta\right)}\right), \qquad (3.2.8)
\end{aligned}$$

onde foi utilizado que $f\left(x^*\right) = 0$. A partir de (3.2.8), conclui-se que se $f'\left(x^*\right) \neq 0$, então o limite $\lim_{n\to+\infty} \dfrac{\left|x^{(n+1)} - x^*\right|}{\left|x^{(n)} - x^*\right|^2}$ existe e vale

$$\lim_{n\to+\infty} \frac{\left|x^{(n+1)} - x^*\right|}{\left|x^{(n)} - x^*\right|^2} = \frac{1}{2}\left|\frac{f''\left(x^*\right)}{f'\left(x^*\right)}\right|$$

pois η,$\xi \to x^*$ quando $n \to +\infty$. Se no entanto, o zero for de multiplicidade maior, ou seja, se

$f'(x^*) = 0$ mas $f''(x^*) \neq 0$, então o desenvolvimento é dado por

$$\begin{aligned} x^{(n+1)} - x^* &= \phi\left(x^{(n)}\right) - \phi(x^*) \\ &= x^{(n)} - \frac{f\left(x^{(n)}\right)}{f'\left(x^{(n)}\right)} - x^* \\ &= x^{(n)} - x^* - \frac{\frac{1}{2}\left(x^{(n)} - x^*\right)^2 f''(\xi)}{\left(x^{(n)} - x^*\right) f''(\eta)} \\ &= x^{(n)} - x^* - \frac{\frac{1}{2}\left(x^{(n)} - x^*\right) f''(\xi)}{f''(\eta)} \\ &= \left(x^{(n)} - x^*\right)\left(1 - \frac{1}{2}\frac{f''(\xi)}{f''(\eta)}\right), \end{aligned}$$

ou seja, a convergência é linear com constante $\frac{1}{2}$,

$$\lim_{n\to+\infty} \frac{\left|x^{(n+1)} - x^*\right|}{\left|x^{(n)} - x^*\right|} = \frac{1}{2}.$$

3.3 Métodos de múltiplos pontos

3.3.1 Método da secante

O método da secante é similar ao método da falsa posição, diferem entre si pelo fato de que no método da secante não há divisão e escolha de intervalos, a sequência de aproximações é calculada a partir das duas últimas aproximações e portanto, devemos iniciar com duas aproximações para a solução. Ao contrário do método da falsa posição, não há necessidade de que a solução esteja entre as duas aproximações iniciais.

A sequência é montada a partir da regra para iteração[6]

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

De maneira semelhante à que ocorre nos métodos de ponto fixo, para que ocorra convergência, em geral, as duas primeiras aproximações devem estar em uma vizinhança suficientemente próxima da solução.

É possível demonstrar que se f for duas vezes continuamente diferenciável e $f'(x^*) \neq 0$, então

[6] É comum utilizar as seguintes variações para minimizar os efeitos de arredondamento:

$$x^{(n+1)} = \begin{cases} x^{(n)} - \dfrac{\left(x^{(n)} - x^{(n-1)}\right)\frac{f\left(x^{(n)}\right)}{f\left(x^{(n-1)}\right)}}{1 - \frac{f\left(x^{(n)}\right)}{f\left(x^{(n-1)}\right)}}, & \text{se } \left|f\left(x^{(n)}\right)\right| < \left|f\left(x^{(n-1)}\right)\right| \\ x^{(n)} - \dfrac{\left(x^{(n)} - x^{(n-1)}\right)}{1 - \frac{f\left(x^{(n-1)}\right)}{f\left(x^{(n)}\right)}}, & \text{se } \left|f\left(x^{(n-1)}\right)\right| < \left|f\left(x^{(n)}\right)\right| \end{cases}$$

existe um constante K tal que

$$\lim_{n\to\infty} \frac{|x^{(n+1)} - x^*|}{|x^{(n)} - x^*|^{\varphi}} = K,$$

onde $\varphi = \dfrac{1+\sqrt{5}}{2} \approx 1{,}618$. Ou seja, apesar de ser mais lenta que no método Newton-Raphson, a convergência é mais rápida que a convergência linear de alguns métodos de ponto fixo.

iteração n	x_n
1	0,544221...
2	0,568826...
3	0,567150...
4	0,567143...

Tabela 3.5: Tabela das primeiras iterações para o método da secante para $f(x) = x - e^{-x}$, com aproximações iniciais $x^{(0)} = 0{,}9$ e $x^{(1)} = 1{,}0$.

a sequência converge para a solução exata até o sexto dígito em quatro iterações. Se utilizarmos $x^{(0)} = 0{,}5$ e $x^{(1)} = 1{,}0$ como primeiras aproximações obteríamos o mesmo resultado após três iterações.

3.4 Raízes de polinômios

As equações não lineares constituídas por polinômios de grau $n \in \mathbb{N}$ com coeficientes complexos $a_n, a_{n-1}, \ldots a_0$:

$$p(x) := a_n x^n + a_{n-1} x^{n-1} + \ldots a_1 x + a_0 = 0, \tag{3.4.1}$$

dispõe de vários métodos para determinar aproximações para suas raízes. Esses métodos foram desenvolvidos a partir da própria estrutura matemática dos polinômios.

O teorema fundamental da álgebra garante que a equação (3.4.1) possui n soluções (denominadas raízes de $p(x)$) no plano complexo, x_1^*, x_2^*, ..., $x_n^* \in \mathbb{C}$. Portanto, $p(x)$ pode ser reescrito como

$$p(x) \equiv (x - x_1^*)(x - x_2^*) \ldots (x - x_n^*).$$

A partir dessa estrutura é possível desenvolver vários métodos específicos para determinar as raízes.

O seguinte teorema relaciona a localização das raízes de um polinômio no plano complexo aos seus coeficientes:

Teorema 3.4.1
Seja $x*$ qualquer raiz do polinômio

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0.$$

Então

$$\begin{aligned}
|x*| &\leq \max\left\{\left|\frac{a_0}{a_n}\right|, 1+\left|\frac{a_1}{a_n}\right|, 1+\left|\frac{a_2}{a_n}\right|, \ldots, 1+\left|\frac{a_{n-1}}{a_n}\right|\right\}, \\
|x*| &\leq \max\left\{1, \sum_{j=0}^{n-1}\left|\frac{a_j}{a_n}\right|\right\}, \\
|x*| &\leq 2\max\left\{\left|\frac{a_{n-1}}{a_n}\right|, \left|\frac{a_{n-2}}{a_n}\right|^{\frac{1}{2}}, \left|\frac{a_{n-3}}{a_n}\right|^{\frac{1}{3}}, \ldots, \left|\frac{a_0}{a_n}\right|^{\frac{1}{n}}\right\}, \\
|x*| &\leq \sum_{j=0}^{n-1}\left|\frac{a_j}{a_{j+1}}\right|.
\end{aligned}$$

A partir desse teorema podemos determinar um disco no plano complexo, com centro na origem, a partir do qual escolhemos aproximações iniciais que serão as entradas de métodos iterativos. Por exemplo, podemos utilizar um dos ponto desse disco como aproximação inicial no método de Newton-Raphson.

3.5 Newton–Raphson modificado

Uma vez definida a região em que as raízes se encontram no plano complexo, escolhemos um ponto nessa região $x^{(0)}$ como aproximação inicial e utilizamos o método de Newton-Raphson usual :

$$x^{(j+1)} = x^{(j)} - \frac{p\left(x^{(j)}\right)}{p'\left(x^{(j)}\right)}.$$

Uma vez determinada a aproximação para a primeira raiz x_1, temos que as demais raízes são também solução do novo polinômio $p_1(x) = \dfrac{p(x)}{x - x_1}$, pois de acordo com o teorema fundamental da álgebra, se x_1, x_2, ..., x_n são raízes de $p(x) = \prod_{j=1}^{n}(x - x_j)$, então, por construção o polinômio $p_1(x)$ possuirá também raízes x_2, x_3, ..., x_n. Portanto para determinar as raízes seguintes utilizamos a regra de Newton-Raphson para a equação $p_1(x) = 0$:

$$x^{(j+1)} = x^{(j)} - \frac{p_1\left(x^{(j)}\right)}{p_1'\left(x^{(j)}\right)}.$$

Porém dado que

$$\frac{p_1(x)}{p_1'(x)} = \frac{p\left(x^{(j)}\right)}{p'\left(x^{(j)}\right) - p\left(x^{(j)}\right)\dfrac{1}{x - x_1}}$$

podemos montar a regra a partir do polinômio original $p(x)$ e da raiz conhecida x_1:

$$x^{(j+1)} = x^{(j)} - \frac{p\left(x^{(j)}\right)}{p'\left(x^{(j)}\right) - p\left(x^{(j)}\right)\dfrac{1}{x - x_1}}$$

e de uma outra aproximação inicial para determinar a nova raiz x_2. Esse procedimento pode ser repetido sucessivamente e uma vez que conheçamos k raízes de $p(x)$, a $k+1$-ésima raiz pode ser determinada através da iteração

$$x^{(j+1)} = x^{(j)} - \frac{p\left(x^{(j)}\right)}{p'\left(x^{(j)}\right) - p\left(x^{(j)}\right)\sum_{m=1}^{k}\frac{1}{x - x_m}}.$$

Naturalmente, as propriedades de convergência desse método são as mesmas do método de Newton-Raphson, ou seja, se a raiz for simples, a convergência será quadrática, se for múltipla a convergência será linear apenas.

3.6 Sistemas de Equações não lineares (método de Newton–Raphson)

Seja um domínio $U \subseteq \mathbb{R}^n$ e $F : \mathbb{U} \to \mathbb{R}^n$ um função contínua e diferenciável. O sistema de equações

$$F(\mathbf{x}^*) = \mathbf{0} \in \mathbb{R}^n, \tag{3.6.1}$$

onde a solução, $\mathbf{x}^* \in \mathbb{R}^n$, é representada por um vetor de componentes $x_1^*, x_2^*, \ldots, x_n^*$ e F possui componentes $F_1(\mathbf{x}), F_2(\mathbf{x}), \ldots, F_n(\mathbf{x}) \in \mathbb{R}$, cada uma delas é uma função contínua e diferenciável de n variáveis. Cada equação $F_i(\mathbf{x}) = 0$ de (3.6.1) define uma hipersuperfície de dimensão $n-1$ contida em $U \subseteq \mathbb{R}^n$. O conjunto solução é formado pela intersecção dessas superfícies.

A natureza do problema é mais complexa, pois conforme aumentamos a dimensão do sistema, o comportamento das soluções é potencialmente mais rico. Como exemplo, vamos considerar o seguinte caso em dimensão dois. Seja $F(x,y) = (F_1(x,y), F_2(x,y))$ onde $F_1(x,y) = \cos(x)\cos(y) - 0.1$ e $F_2(x,y) = \text{sen}(x)\text{sen}(y) - 0.5$. A figura 3.6.1 contém o gráfico dessas funções.

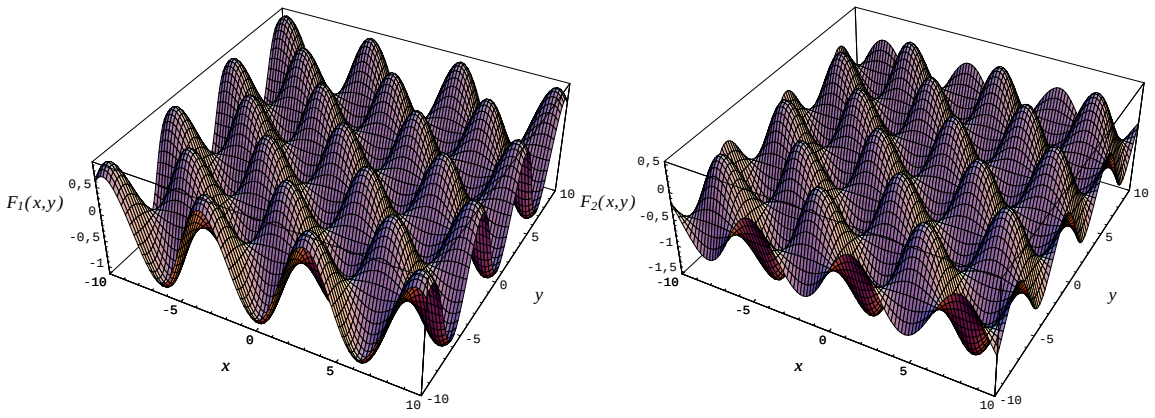


Figura 3.6.1: Gráficos das funções $F_1(x,y)$e $F_2(x,y)$

Cada equação $F_1(x,y) = 0$ e $F_2(x,y) = 0$ determina um conjunto de curvas (superfícies contidas em um espaço de dimensão dois) no plano $x \times y$. As soluções de $F(\mathbf{x}) = \mathbf{0} \in \mathbb{R}^2$, ou dito de outra forma, do sistema

$$\begin{cases} F_1(x,y) = 0 \\ F_2(x,y) = 0 \end{cases}$$

são as intersecções desses dois conjuntos de superfícies. A figura 3.6.2 contém o conjunto de pontos no plano $x \times y$ que satisfazem as equações e o sistema.

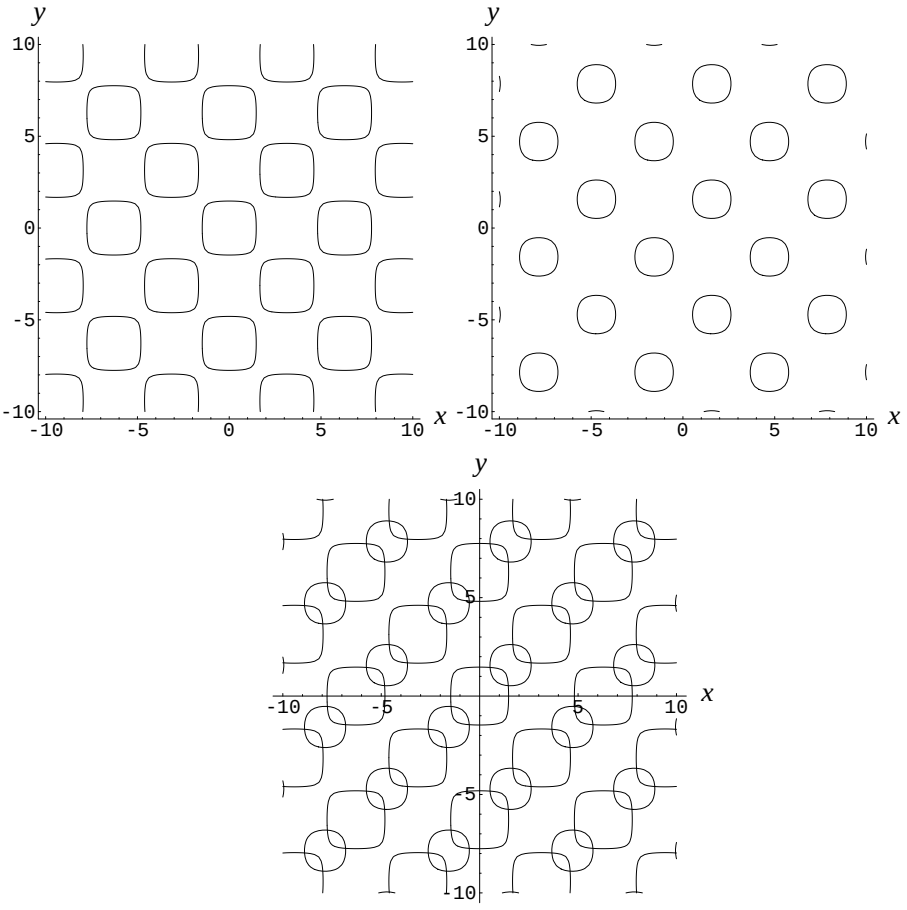


Figura 3.6.2: Acima à esquerda: curvas formadas pelos conjuntos de pontos que satisfazem $\cos(x)\cos(y) = 0{,}1$. Acima à direita: curvas formadas pelos conjuntos de pontos que satisfazem $\text{sen}(x)\text{sen}(y) = 0{,}5$. Abaixo: sobreposição das curvas (superfícies); os pontos de intersecção das curvas são as soluções do sistema.

Existe um número menor de métodos disponíveis para determinar uma solução aproximada para (3.6.1). Vamos estudar apenas a extensão do método Newton-Raphson para sistemas de equações não lineares.

Se $A(\mathbf{x})$ for uma matriz $n \times n$ não singular (determinante diferente de zero) em alguma vizinhança da solução $\mathbf{x}^*$ da equação (3.6.1), então $\mathbf{x}^*$ também é solução da equação

$$\Phi(\mathbf{x}^*) = \mathbf{x}^*,$$

onde $\Phi(\mathbf{x}) := \mathbf{x} + A(\mathbf{x})F(\mathbf{x})$. A função Φ permite a construção de uma sequência de aproximações $\left\{\mathbf{x}^{(j)}\right\}_{j=0}^{\infty}$ a partir da regra

$$\mathbf{x}^{(j+1)} = \Phi\left(\mathbf{x}^{(j)}\right)$$

e de uma aproximação inicial $\mathbf{x}^{(0)}$. A questão da convergência dos elementos da sequência para a solução é tratada por uma forma mais geral do teorema do ponto fixo que estudamos na seção 3.2. O teorema garante que se $\mathbf{x}^{(0)}$ for suficientemente próximo da solução então existe um $0 \leq K < \infty$

$$\lim_{j\to\infty} \frac{\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{(j)} - \mathbf{x}^*\right\|} = K,$$

onde a notação $\|\cdot\|$ indica uma norma (medida) para os vetores em dimensão n.

Também de modo análogo ao estudado na seção 3.2, o método de Newton-Raphson consiste em determinar uma função Φ tal que a convergência seja quadrática, i. e., Φ deve ser tal que os elementos da sequência $\left\{\mathbf{x}^{(j)}\right\}_{j=0}^{\infty}$ possuam a seguinte convergência[7]:

$$\lim_{j\to\infty} \frac{\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{(j)} - \mathbf{x}^*\right\|^2} = \tilde{K}$$

para um $0 \leq \tilde{K} < \infty$. A função Φque garante esse comportamento é $\Phi(\mathbf{x}) := \mathbf{x} - J^{-1}(\mathbf{x})F(\mathbf{x})$, onde $J^{-1}(x)$ é a inversa da matriz jacobiana da transformação (ou função) $F(\mathbf{x})$. As componentes i, j da matriz jacobiana são dadas por

$$(J(\mathbf{x}))_{i,j} = \frac{\partial F_i}{\partial x_j}(\mathbf{x})$$

para $F(\mathbf{x}) = (F_1(\mathbf{x}), F_{12}(\mathbf{x}), \ldots, F_n(\mathbf{x}))$, onde cada i-ésima componente é da forma $F_i(\mathbf{x}) = F_i(x_1, x_2, \ldots, x_n)$.

O método consiste na construção da sequência de aproximações $\left\{\mathbf{x}^{(j)}\right\}_{j=0}^{\infty}$ a partir de uma aproximação inicial $\mathbf{x}^{(0)}$ e da regra

$$\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} - J^{-1}(\mathbf{x}^{(j)})F(\mathbf{x}^{(j)})$$

para $j \geq 0$.

Exemplo 19: A equação $z^* - e^{-z^*} = 0$ possui apenas uma solução real, dada pela função W de Lambert, $z^* = W(1) = 0{,}567143290\ldots$ Porém, essa mesma equação possui infinitas soluções no plano complexo. É possível determinar aproximações para as soluções complexas a partir de uma aproximação inicial complexa e do método de Newton-Raphson usual.

De acordo com o método, utilizamos a sequência

$$z^{(j+1)} = \frac{1 + z^{(j)}}{1 + e^{z^{(j)}}}$$

Vamos representar as partes real e imaginária do complexo z por x e y, respectivamente, ou seja, $z = x + iy$. Então, segundo a fórmula de Euler, a regra anterior implica as seguintes regras para as partes real e imaginária das aproximações

$$\begin{aligned} x^{(j+1)} + iy^{(j+1)} &= \frac{1 + x^{(j)} + iy^{(j)}}{1 + e^{x^{(j)} + iy^{(j)}}} \\ &= \frac{1 + x^{(j)} + iy^{(j)}}{1 + e^{x^{(j)}}\left(\cos(y^{(j)}) + i\text{sen}(y^{(j)})\right)}, \end{aligned}$$

[7] O que necessariamente implica $\lim_{j\to\infty} \dfrac{\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^*\right\|}{\|\mathbf{x}^{(j)} - \mathbf{x}^*\|} = 0$.

finalmente

$$x^{(j+1)} = \frac{\left(1 + x^{(j)}\right)\left(e^{(j)} + \cos(y^{(j)})\right) + y^{(j)}\text{sen}(y^{(j)})}{2\left(\cos(y^{(j)}) + \cosh(x^{(j)})\right)} \tag{3.6.2}$$

e

$$y^{(j+1)} = \frac{y^{(j)}\left(e^{-x^{(j)}} + \cos(y^{(j)})\right) - \left(1 + x^{(j)}\right)\text{sen}(y^{(j)})}{2\left(\cos(y^{(j)}) + \cosh(x^{(j)})\right)}. \tag{3.6.3}$$

De maneira alternativa, poderíamos considerar o problema bidimensional como uma equação que leva um número complexo (representado por um vetor com duas componentes reais) em outro complexo :

$$x^* + iy^* - e^{-x^* - iy^*} = 0 = 0 + 0i.$$

De acordo com a fórmula de Euler, termo $e^{-iy^*} = \cos(y^*) - i\text{sen}(y^*)$, portanto a equação anterior assume a forma

$$x^* - e^{-x^*}\cos(y^*) + i\left(y + e^{-x^*}\text{sen}(y^*)\right) = 0 + 0i.$$

Ou seja,

$$F(x^*, y^*) = (F_1(x^*, y^*), F_2(x^*, y^*)) = (0, 0),$$

onde

$$F_1(x, y) = x - e^{-x}\cos(y)$$

e

$$F_2(x, y) = y + e^{-x}\text{sen}(y).$$

Nesse caso a matriz jacobiana assume a forma

$$J(x, y) = \begin{pmatrix} 1 + e^{-x}\cos(y) & e^{-x}\text{sen}(y) \\ -e^{-x}\text{sen}(y) & 1 + e^{-x}\cos(y) \end{pmatrix},$$

cujo determinante é

$$\det J(x, y) = 1 + 2e^{-x}\cos(y) + e^{-2x} = 2e^{-x}\left(\cos(y) + \cosh(x)\right).$$

Como $\cos(y) \geq -1$, temos que $\det J(x, y) \geq 1 + 2e^{-x} + e^{-2x} = (1 + e^{-x}) \geq 0$ só se anula quando $x = 0$, ou seja, no eixo imaginário. Fora dessa região a matriz jacobiana é não singular e aí o método está bem definido.

A inversa de $J(x, y)$ é dada por

$$J^{-1}(x, y) = \begin{pmatrix} \dfrac{e^x + \cos(y)}{2\left(\cos(y) + \cosh(x)\right)} & -\dfrac{\text{sen}(y)}{2\left(\cos(y) + \cosh(x)\right)} \\ \dfrac{\text{sen}(y)}{2\left(\cos(y) + \cosh(x)\right)} & \dfrac{e^x + \cos(y)}{2\left(\cos(y) + \cosh(x)\right)} \end{pmatrix}.$$

De onde podemos verificar que a regra

$$\left(x^{(j+1)}, y^{(j+1)}\right) = \left(x^{(j)}, y^{(j)}\right) - J^{-1}\left(x^{(j)}, y^{(j)}\right) F\left(x^{(j)}, y^{(j)}\right)$$

é exatamente igual às regras (3.6.2) e (3.6.3).

A figura 3.6.3 ilustra o comportamento da função $|F(x, y)|$, as áreas mais claras representam regiões de valores próximos do zero. As soluções da equação correspondem ao centros dos elipsoides.

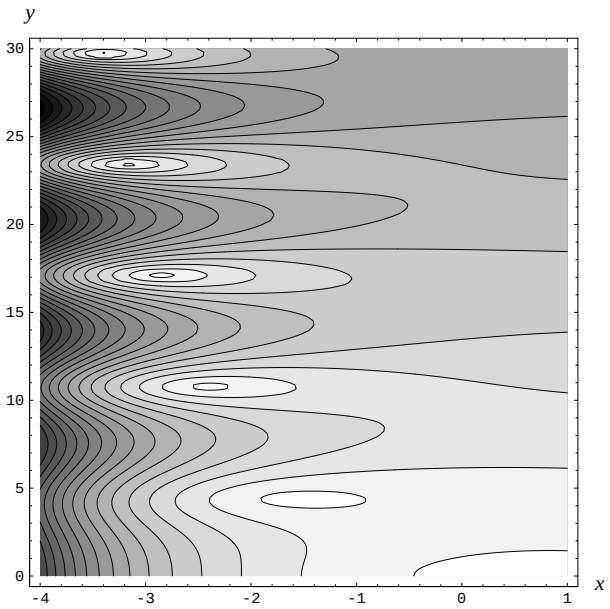


Figura 3.6.3: Curvas de nível da função $|x + iy - e^{-(x+iy)}|$. Regiões mais claras representam valores menores.

A partir do gráfico, escolhemos como valor inicial o complexo $-1{,}0+4{,}0i$ que está próxima da solução complexa com menor valor absoluto para a parte imaginária. Abaixo, segue a tabela com o resultado das iterações com as regras (3.6.2) e (3.6.3):

iteração n	x_n	y_n
1	$-1{,}70175\ldots$	$4{,}64257\ldots$
2	$-1{,}52548\ldots$	$4{,}42130\ldots$
3	$-1{,}53277\ldots$	$4{,}37503\ldots$
4	$-1{,}53391\ldots$	$4{,}37518\ldots$
5	$-1{,}53391\ldots$	$4{,}37518\ldots$

Tabela 3.6: Tabela das primeiras iterações do método de Newton-Raphson para $f(x + iy) = x + iy - e^{-(x+iy)}$, com aproximações iniciais $x_0 = -1{,}0$ e $y_0 = 4{,}0$.

3.7 Exemplos comentados

Problema 1

Uma estação de bombeamento é responsável por um fluxo $\phi(r)$ de fluido (em l/s) , onde r é a velocidade de rotação do motor da bomba (em 10^3rad/s). Para manter o motor com uma rotação r , a estação necessita de uma potência $P(r)$ (em kW). O objetivo é determinar o intervalo de valores de r, (r_{min}, r_{max}), que correspondem à faixa superior a 90% do rendimento máximo (em l/J) se

$$\phi(r) = 500 \tanh(0{,}87r) \qquad \text{e} \qquad P(r) = 2 + 3{,}078^r.$$

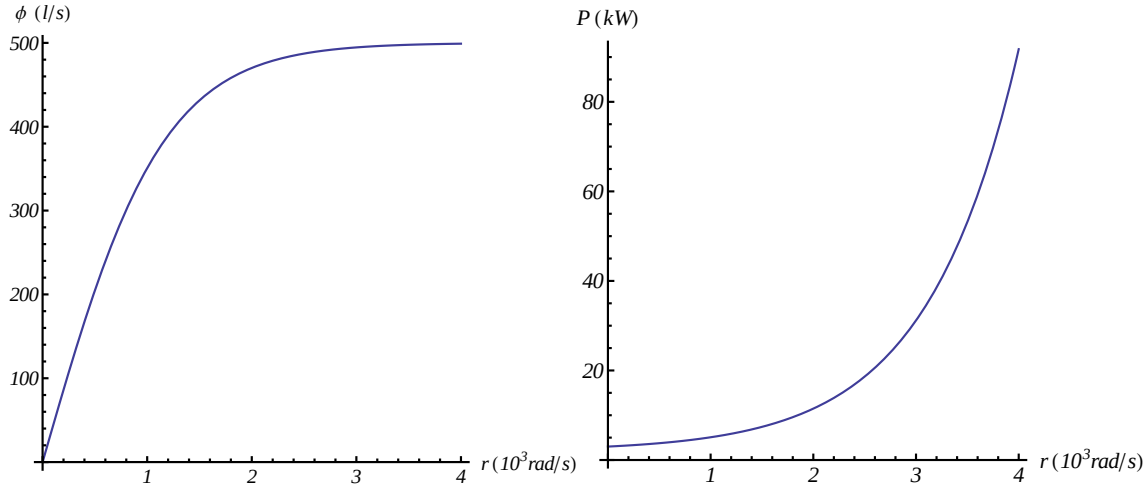


Figura 3.7.1: Comportamento das funções ϕ e P.

A partir dos gráficos, é possível notar que conforme a rotação do motor é aumentada, o gasto energético cresce exponencialmente enquanto que a quantidade de fluido bombeado atinge um ponto de saturação. É natural esperar que exista uma situação ótima, na qual há um máximo transporte de fluido por unidade de energia.

Se a velocidade de rotação é mantida constante, em um intervalo de tempo Δt, a estação bombeia um volume de fluido $\phi(r)\Delta t$. Nesse mesmo intervalo de tempo, ela gasta uma quantidade de energia $P(r)\Delta t$. Assim o rendimento em litros por joules é dado por

$$\nu(r) := \frac{\phi(r)\Delta t}{P(r)\Delta t} = \frac{\phi(r)}{P(r)},$$

ou seja,

$$\nu(r) = \frac{500 \tanh(0{,}87r)}{2 + 3{,}078^r}.$$

O gráfico da função ν permite notar que a exigência de trabalho acima de 90% da eficiência máxima corresponde a um intervalo da valores (r_{min}, r_{max}) para a velocidade de rotação do motor.

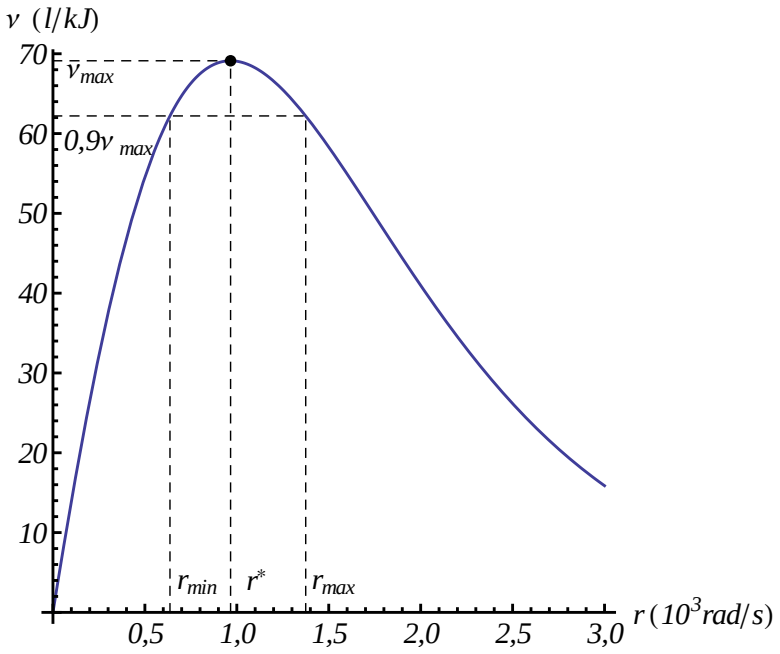


Figura 3.7.2: Gráfico para o rendimento em função da velocidade de rotação.

No valor r^*, onde ν é máximo, temos $\nu'(r^*) = 0$. A primeira tarefa consiste em determinar o zero r^* de ν'

$$\nu'(r) = 500\left(\frac{0{,}87}{2+3{,}078^r}\operatorname{sech}^2(0{,}87r) - \ln(3{,}078)\frac{3{,}078^r\tanh(0{,}87r)}{(2+3{,}078^r)^2}\right).$$

A máxima eficiência ν_{max} é dada por

$$\nu_{max} = \nu(r^*).$$

Assim, uma vez determinado r^*, os valores r_{min} e r_{max} são raízes da equação

$$\nu(x) - 0{,}9\nu_{max} = 0.$$

As seguintes instruções no Scilab fornecem o resultado numérico.

```
// Definição da função ni(r)
function z=ni(r)
  z=500*tanh(0.87*r)/(2+3.078^r);
endfunction

// Valores de r utilizados para desenhar o gráfico.
rvar=linspace(0.5,1.5,100);
// Gráfico de ni
fplot2d(rvar,ni);
xgrid;

// Determinação de r* que corresponde ao máximo de ni'
// Definição da derivada de ni
```

```
function z=d_ni(r)
  zl=500*(0.87*sech(0.87*r)^2/(2+3.078^r);
  z2=log(3.078)*3.078^r*tanh(0.87*r)/((2+3.078^r)^2));
  z=zl-z2;
endfunction

// Definição da derivada aproximada de ni
// através da instrução derivative. (Forma alternativa)
function z=d_ni2(r)
  z=derivative(ni,r);
endfunction

// Gráfico de ni'
scf(); // abre uma nova janela gráfica
fplot2d(rvar,d_ni);
xgrid;

// Determina r*
// Aproximação inicial é 0.96
r_estrela=zero_newraph(d_ni,0.96,l00);
// Forma alternativa, a partir da derivada numérica.
r_estrela2=zero_newraph(d_ni2,0.96,l00);
ni_max=ni(r_estrela);

// Determinar r_min e r_max
// Esses valore são zeros da função f
function z=f(x)
  z=ni(x)-0.9*ni_max;
endfunction

// Aproximação inicial para r_min é 0.64
r_min=zero_newraph(f,0.64,l00);
// Aproximação inicial para r_max é l.38
r_max=zero_newraph(f,l.38,l00);
```

Problema 2

Considere o circuito elétrico descrito pela figura seguinte

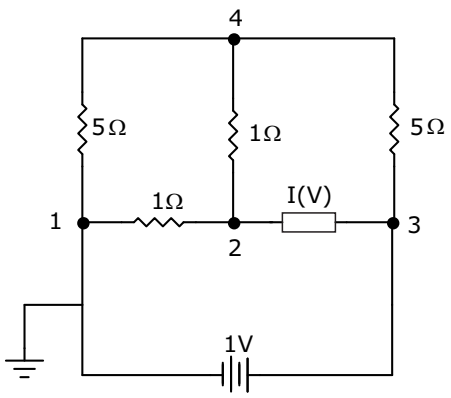


Figura 3.7.3: Circuito elétrico. O elemento não linear possui um relação entre tensão e corrente dada pela equação (3.7.1).

Este circuito é semelhante ao estudado na seção 2.4. A tensão no nó 1 é fixa em 0V e a tensão no nó 3 em 1V. As variáveis deste sistema são sete: V_2, V_4, $I_{1,2}$, $I_{2,3}$, $I_{3,4}$, $I_{1,4}$ e $I_{2,4}$. As equações são dadas pelas leis de Ohm (2.4.5), Kirchoff para correntes (2.4.6) e pela equação que relaciona tensão e corrente no elemento não linear,

$$I_{2,3} = (V_2 - V_3)^3 - 2\,(V_2 - V_3)^2 + 1{,}04\,(V_2 - V_3)\,. \tag{3.7.1}$$

A seguinte escolha para o ordenamento das variáveis será adotada

$$x_1 := V_2,\ x_2 := V_4,\ x_3 := I_{1,2},\ x_4 = I_{2,3},\ x_5 = I_{3,4},\ x_6 = I_{1,4},\ x_7 = I_{2,4}.$$

As equações serão ordenadas da seguinte forma:

- 1ª equação: lei de Ohm para o resistor ligado aos nós 1 e 2.
- 2ª equação: dada por (3.7.1).
- 3ª equação: lei de Ohm para o resistor ligado aos nós 3 e 4.
- 4ª equação: lei de Ohm para o resistor ligado aos nós 2 e 4.
- 5ª equação: lei de Ohm para o resistor ligado aos nós 1 e 4.
- 6ª equação: lei de Kirchoff para o nó 2.
- 7ª equação: lei de Kirchoff para o nó 4.

De acordo com o ordenamento escolhido, o sistema de equações é dado por

$$\begin{cases} -x_1 - x_3 & = 0 \\ -(x_1-1)^3 + 2(x_1-1)^2 - 1{,}04(x_1-1) + x_4 & = 0 \\ -x_2 - 5x_5 + 1 & = 0 \\ x_1 - x_2 - x_7 & = 0 \\ -x_2 - 5x_6 & = 0 \\ x_3 - x_4 - x_7 & = 0 \\ x_5 + x_6 + x_7 & = 0 \end{cases} \tag{3.7.2}$$

Uma forma de se obter uma aproximação inicial, $x^{(0)}$, para a solução de (3.7.2) consiste em substituir a segunda equação por uma versão livre de termos não lineares

$$-1{,}04(x_1 - 1) + x_4 = 0.$$

Nesse caso, a aproximação inicial é solução do sistema de equações lineares

$$Ax^{(0)} = b,$$

onde

$$A = \begin{pmatrix} -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1{,}04 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -5 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & -2 & 0 & 0 & 0 & -5 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \quad \text{e} \quad b = \begin{pmatrix} 0 \\ -1{,}04 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

As seguintes instruções no Scilab fornecem o resultado numérico.

```
// Voltagem da fonte
V=1
// Aproximação inicial
A=zeros(7,7);
A(1,1)=-1; A(1,3)=-1;
A(2,1)=-1.04; A(2,4)=1;
A(3,2)=-1; A(3,5)=-5;
A(4,1)=1; A(4,2)=-1; A(4,7)=-1;
A(5,2)=-1; A(5,6)=-5;
A(6,3)=1; A(6,4)=-1; A(6,7)=-1;
A(7,5)=1; A(7,6)=1; A(7,7)=1;
b=zeros(7,1);
b(2)=-1.04*V;
b(3)=-V;
```

```
x0=A\b;

// Definição da função F.
function z=F(x)
  z(1)=-x(1) - x(3)
  z(2)=-(x(1)-V)^3 + 2*(x(1)-V)^2 - 1.04*(x(1)-V) + x(4)
  z(3)=-x(2) - 5*x(5) + V
  z(4)= x(1) - x(2) - x(7)
  z(5)=-x(2) - 5*x(6)
  z(6)= x(3) - x(4) - x(7)
  z(7)= x(5) + x(6) + x(7)
endfunction

// Definição da jacobiana dF.
function z=dF(x)
  z=A
  z(2,1)=-3*(x(1)-V)^2 + 4*(x(1)-V) - 1.04
endfunction

// Solução com estimativa para tolerância de 1e-10 na norma 1.
sol=fsolve_nr(x0,F,dF);

// Corrente que deixa a fonte
I=sol(5)-sol(4);
// Corrente que deixa a fonte na aproximação inicial
// (aproximação linear)
I0=x0(5)-x0(4);
```

A corrente que deixa a fonte em direção ao nó 3 é obtida a partir da solução do problema e da Lei de Kirchoff,

$$I_{fonte,3} + I_{2,3} + I_{4,3} = 0.$$

Como $I_{4,3} = -I_{3,4}$, a equação anterior pode ser reescrita como

$$\begin{aligned} I_{fone,3} &= -I_{2,3} + I_{3,4} \\ &= -x_4 + x_5. \end{aligned}$$

A partir da solução obtida numericamente temos

$$I_{fonte,3} \approx 760{,}07mA.$$

3.8 Exercícios

1) Seja a equação não linear

$$x - e^{-x} = 0.$$

A solução é dada em termos da função W de Lambert, $x^* = W(1) \approx 0{,}567143290\ldots$ Se utilizarmos o método da bissecção e o intervalo inicial $(0, 1)$ serão necessárias 20 iterações para obter um resultado com 6 casas decimais exatas. Utilizando o mesmo intervalo inicial mas com o método da falsa posição serão necessárias apenas 8 iterações para obter um resultado com a mesma exatidão.

Se no entanto, o intervalo inicial for $(-10, 10)$ serão necessárias 22029 iterações no método da falsa posição enquanto que no método da bissecção serão necessárias apenas 24 iterações. Como você explicaria essa diferença?

2) Encontre as duas soluções reais da equação

$$-x + e^x - 3 = 0$$

com seis dígitos exatos.

3) As seguintes equações possuem uma raiz real positiva igual a $\dfrac{3}{2}$.

$$x^4 - 3{,}5\,x^3 + 2{,}25\,x^2 + 3{,}375\,x - 3{,}375 = 0$$

$$x^4 + 1{,}5\,x^3 - 1{,}5\,x^2 - 3{,}5\,x - 1{,}5 = 0$$

Utilize o método de Newton-Raphson com algumas aproximações iniciais diferentes para encontrar essa raiz. O que você pode notar?

4) Utilize os métodos de Newton-Raphson e da secante para determinar os seis primeiros dígitos da primeira solução real positiva da equação

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

5) Utilize os métodos de Newton-Raphson e da secante para determinar os seis primeiros dígitos das duas soluções reais e positivas da equação

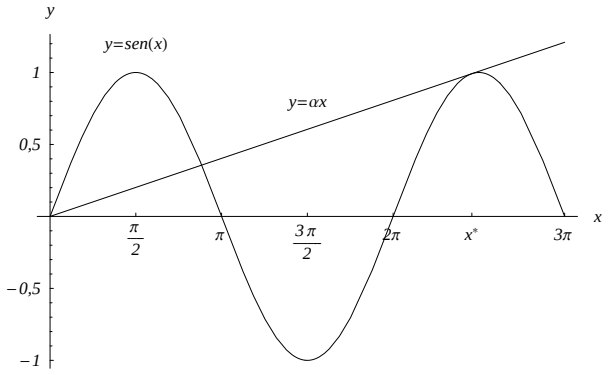
$$x^x - 0{,}8 = 0.$$

6) Determine os seis primeiros dígitos das três soluções reais e positivas da equação

$$\cos(x) = 0{,}02\,x^2.$$

7) A partir do método de Newton-Raphson, desenvolva um algoritmo para calcular a raiz quadrada de um ponto flutuante x no menor número de passos possível. Sugestão: utilize a base 2, represente $x = \text{mantissa}(x) \times 2^{\text{expoente}(x)}$ e trabalhe com uma escolha adequada de aproximações iniciais, como por exemplo retas ajustadas para cada domínio de valores de x.

8) A partir do gráfico abaixo, determine uma aproximação para o ponto x^*em que a reta $y = \alpha x$ tangencia a curva $y = \text{sen}(x)$. Utilize o método de Newton-Raphson e obtenha 5 dígitos exatos.



9) Determine as soluções mais próximas do eixo $(0, 0)$ para o sistema de equações não lineares da figura 3.6.2. Apresente o resultado com 5 dígitos exatos.

10) Determine todas as raízes (e suas respectivas multiplicidades) do polinômio

$$x^6 - x^5 + x^4 - \frac{3}{4}x^3 - \frac{1}{16}x^2 + \frac{1}{4}x - \frac{1}{16}$$

com 5 dígitos exatos.

11) Utilize o método da secante para determinar a melhor aproximação com 6 dígitos de precisão para o valor $x^* > 0$ que corresponde ao primeiro mínimo da função $\dfrac{\text{sen}\, 3x}{x}$.

12) A partir do método Newton-Raphson é possível montar uma relação de recorrência para obter a raiz de índice $k = 2, 3, \ldots$ de um número $y > 0$, $\sqrt[k]{y}$, através de operações elementares $(+, -, \times \text{ e } \div)$. Obtenha a relação de recorrência que aproxima $\sqrt[8]{10}$. *Sugestão*: note que a raiz é solução de $x^8 = 10$.

13) A relação entre pressão P (em unidades $\text{Pa} = \text{N/m}^2$), temperatura T (K) e volume específico $v(\text{m}^3/\text{Kg})$ em um gás não ideal é aproximada pela equação de estado de van der Waals

$$\left(P + \frac{a}{v^2}\right)(v - b) = RT,$$

onde $R = 461{,}495\,\text{J}/\,(\text{Kg K})$, $a = 1703{,}28\,\text{Pa m}^6/\text{Kg}^2$ e $b = 0{,}00169099\,\text{m}^3/\text{Kg}$. Construa uma tabela com os valores da densidade desse gás (o recíproco do volume específico) para valores da temperatura entre $500K$ e $1000K$ com espaçamento de $100K$ quando $P = 10^5\text{Pa}$. Os valores tabelados devem conter cinco dígitos.

14) Considere uma bomba movida a energia elétrica para a qual a relação "fluxo"× "velocidade de rotação" é dada pela função

$$V(\omega) = 0{,}7 \tanh\left(1{,}1\omega^{1{,}37}\right) + 0{,}3 \tanh\left(\omega^{3{,}87}\right)$$

e a relação entre "potência dissipada"×"velocidade de rotação" é dada pela função

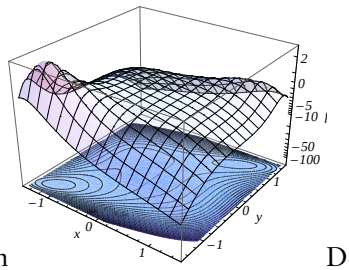
$$P(\omega) = \omega^{1{,}1} \exp\left(0{,}8\sqrt{\omega}\right).$$

A razão $\dfrac{V(\omega)}{P(\omega)}$, determinada taxa de transporte, fornece informação sobre a quantidade de matéria transportada por unidade de energia para uma velocidade de rotação ω. Dado que a taxa máxima ocorre em $\omega_{maxtrans} \approx 0{,}715574$, determine com uma precisão de seis dígitos o intervalo de valores para a rotação no qual a bomba opera em uma faixa superior a 85% da taxa máxima de transporte.

15) Considere um sinal cuja dependência no tempo t possui a forma

$$\frac{1}{2}\left(1 + \cos\left(5t\right)\right)\cos^2\left(\pi t\right).$$

Determine com precisão de seis dígitos, por quanto tempo esse sinal é superior a 0,4 no intervalo $t \in [0, 5]$.



16) [l]4cm Determine uma aproximação para o ponto (x^*, y^*) no qual a função $f : \mathbb{R}^2 \to \mathbb{R}$,
$f(x, y) = -x^4 - y^6 + 3xy^3 - x$, atinge seu valor máximo.

17) Considere um investimento financeiro que rende mensalmente um percentual $r > 0$ sobre o montante investido. Se um investidor aplica mensalmente uma quantia Q ao longo de n meses e nos m meses seguintes subtrai mensalmente uma quantia R, o montante que sobra após o m-ésimo resgate será

$$\left(\frac{(1+r)\left((1+r)^n - 1\right)Q - R}{r}\right)(1+r)^{m-1} + \frac{R}{r}.$$

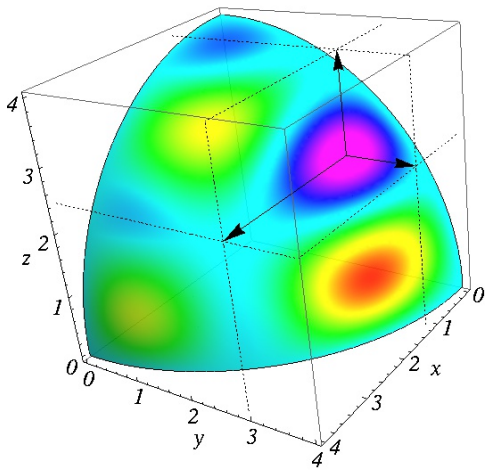
Supondo que um investidor realiza aplicações mensais de de R\$1.500, 00 por 25 anos (300 meses), determine uma aproximação com seis dígitos para o menor valor que o percentual r deve assumir de modo que o montante permita a realização de resgates de R\$7.000, 00 por 35 anos (420 meses)?

18) Determine as 4 soluções do sistema

$$\begin{cases} 5\text{sen}\left(x_1^2\right) - 2\cos(x_2) + 1 &= 0 \\ x_2 - 0{,}75\text{sen}\left(x_1\right) + 0{,}8\cos\left(4{,}41x_2\right) &= 0 \end{cases}$$

na região $[-1, 1] \times [-1, 1]$.

19) O problema de se determinar os valores extremos de uma função sujeita a vínculos em seus argumentos pode ser resolvido com o auxílio dos multiplicadores de Lagrange. Seja a função $F\left(x, y, z\right) := \left(\text{sen}\, x\right)\left(\text{sen}\, \sqrt{2}\, y\right)\left(\text{sen}\, \sqrt{3}\, z\right)$, desejamos determinar os valores (x^*, y^*, z^*) no octante $x > 0$, $y > 0$, $z > 0$ para os quais F assume o seu maior valor, dado que $x^2 + y^2 + z^2 = 4^2$. Uma condição necessária para que (x^*, y^*, z^*) seja um ponto de máximo (ou mínimo) é que seja solução do sistema de equações $\dfrac{\partial \Lambda}{\partial x} = \dfrac{\partial \Lambda}{\partial y} = \dfrac{\partial \Lambda}{\partial z} = \dfrac{\partial \Lambda}{\partial \lambda} = 0$, onde
$\Lambda(x, y, z, \lambda) := F(x, y, z) + \lambda G(x, y, x)$ e $G(x, y, z) = x^2 + y^2 + z^2 - 4^2$ representa o vínculo. Nesse caso, (x^*, y^*, z^*) deve ser solução do sistema de equações



$$\begin{cases} (\cos x)\left(\text{sen}\, \sqrt{2}\, y\right)\left(\text{sen}\, \sqrt{3}\, z\right) + 2\lambda x &= 0 \\ \sqrt{2}(\text{sen}\, x)\left(\cos \sqrt{2}\, y\right)\left(\text{sen}\, \sqrt{3}\, z\right) + 2\lambda y &= 0 \\ \sqrt{3}(\text{sen}\, x)\left(\text{sen}\, \sqrt{2}\, y\right)\left(\cos \sqrt{3}\, z\right) + 2\lambda z &= 0 \\ x^2 + y^2 + z^2 - 4^2 &= 0 \end{cases}$$

Através de um código de cores (vermelho para os menores valores até violeta para os maiores valores), o gráfico ao lado apresenta o valor de F sobre a superfície definida pelo vínculo. A partir da solução numérica, determine uma aproximação de seis dígitos para o valor máximo de F sujeito a esses vínculos. (As coordenadas estão indicadas pelas setas).

20) A relação entre custo de produção (em R$/MWh) e potência (em GW) de três usinas elétricas é dada pelas funções C_1, C_2 e C_3, válidas para potências entre 0 e 5GW,

$$\begin{aligned} C_1(p) &= 15 + 15p + 2p^2 + 0{,}8p^3 \\ C_2(p) &= 30 + 10p + p^2 + 0{,}1p^4 \\ C_3(p) &= 40 + 11p + 0{,}1p^2 + p^3 \end{aligned}$$

A partir da técnica do multiplicador de Lagrange, determine o valor mínimo do custo total de produção para uma potência total de 10GW gerada em conjunto pelas três usinas. De acordo com a técnica, devemos encontrar as potências produzidas nas usinas 1, 2 e 3, representadas respectivamente por p_1, p_2 e p_3 que satisfaçam $\dfrac{\partial \Lambda}{\partial p_i} = 0$, para $i = 1,\ 2,\ 3$ e $\dfrac{\partial \Lambda}{\partial \lambda} = 0$, onde

$$\Lambda\left(p_1, p_2, p_3, \lambda\right) := C\left(p_1, p_2, p_3\right) + \lambda\left(p_1 + p_2 + p_3 - 10\right)$$

e C é a função do custo total de produção, $C\left(p_1, p_2, p_3\right) := \sum_{i=1}^{3} C_i(p_i)$. Observação: uma vez escolhidas as aproximações para p_1, p_2 e p_3, qualquer uma das equações $\dfrac{\partial \Lambda}{\partial p_i} = 0$ pode ser utilizada para produzir uma aproximação inicial para λ.

21) A trajetória de uma satélite orbitando a Terra é descrita em coordenadas polares pela equação

$$r(\theta) = A\frac{1-\varepsilon^2}{1+\varepsilon \operatorname{sen}(\theta+\phi)},$$

onde A é o semieixo maior da órbita (em km), ε é a sua excentricidade e ϕ é uma fase. Determine o valor aproximado (com 4 dígitos) do semieixo maior de uma órbita que passa pelos pontos indicados na tabela abaixo:

θ	$-\pi/6$	0	$\pi/6$
r(km)	6870	6728	6615

.

4 Derivação numérica

Nesta seção vamos desenvolver métodos para estimar a derivada de uma função f calculada em um ponto x^*, $f'(x^*)$, a partir de valores conhecidos de f em pontos próximos ao ponto x^*.

Uma possível abordagem para encontrar a derivada em um ponto x^* consiste em determinar uma interpolação polinomial, $p(x)$, a partir dos valores de f em pontos próximos a x^* e então estimar $f'(x^*)$ a partir de $p'(x^*)$. Essa abordagem é a mais indicada quando estamos interessados no valor da derivada para diversos pontos ou quando os pontos utilizados para construir a interpolação p não estão igualmente espaçados.

Na situação em que a função f é conhecida em uma sequência igualmente espaçada de pontos, dispomos de outras técnicas como o cálculo das derivadas a partir de operações de diferença finita.

Antes de dar continuidade, a seguinte definição nos será muito útil:

Definição 4.0.1 (Notação $O(\cdot)$). *A notação $f(x) = O(g(x))$ quando $x \to x_0$ significa que existem constantes positivas ϵ e δ tais que*

$$|f(x)| \leq \delta\,|g(x)|$$

para todo x no intervalo $|x - x_0| \leq \epsilon$.

A mesma notação utilizada na situação $x \to \infty$(ou $-\infty$) significa que existem constantes $\delta > 0$ e $\tilde{x} > 0$ (respectivamente $\tilde{x} < 0$) tal que a mesma desigualdade é válida para todo $x > \tilde{x}$ (respectivamente $x < \tilde{x}$).

A partir da definição podemos concluir, por exemplo, que $\text{sen}(x) = O(x)$ quando $x \to 0$. Da mesma forma $30(\cos(12x) - 1) = O(x^2)$, $e^x \cos(x) = O(1)$ e $10\cos(x) = O(1)$ quando $x \to 0$.

Além disso, a notação é tal que as seguinte propriedades são satisfeitas: dados $a \geq b \geq 0$, se $f(x) = O(x^a)$ e $g(x) = O(x^b)$ quando $x \to 0$, então $f(x)+g(x) = O(x^b)$, $f(x)-g(x) = O(x^b)$, $f(x)/g(x) = O(x^{a-b})$ e $f(x)g(x) = O(x^{a+b})$ quando $x \to 0$.

Aproximação da derivada por diferenças finitas

A partir da definição da função $f'(x)$ através do limite

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

introduzimos a operação de diferença finita [1] $D_{+,h}$, a partir da qual obtemos uma segunda função $g_h \equiv (D_{+,h}f)$:

$$g_h(x) = (D_{+,h}f)(x) = \frac{f(x+h) - f(x)}{h}, \qquad \text{(diferença progressiva)}.$$

No limite recuperamos a função derivada de f, $\lim_{h\to 0} g_h(x) = \lim_{h\to 0} (D_{+,h}f)(x) = f'(x)$. É importante notar que a definição de derivada a partir de limites não é única, podemos definir a mesma função derivada de f a partir de outros limites (e assim, determinar outras operações de diferença finita), por exemplo

$$f'(x) = \lim_{h\to 0} \frac{f(x) - f(x-h)}{h}$$

ou ainda

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x-h)}{2h}.$$

A cada uma dessas definições podemos associar naturalmente uma operação de diferença finita. A partir dos dois últimos limites, associamos as operações $D_{-,h}$ e $D_{0,h}$:

$$(D_{-,h}f)(x) := \frac{f(x) - f(x-h)}{h}, \qquad \text{(diferença regressiva)}$$

$$(D_{0,h}f)(x) := \frac{f(x+h) - f(x-h)}{2h}, \qquad \text{(diferença central)}.$$

As funções que resultam da ação das duas primeiras operações, $D_{+,h}$ e $D_{-,h}$, sobre uma função f podem ser prontamente identificadas, respectivamente, com a derivada da interpolação de uma reta a partir dos pontos $(x, f(x))$, $(x+h, f(x+h))$ no primeiro caso e $(x, f(x))$, $(x-h, f(x-h))$, no segundo. Já a função que resulta da operação $D_{0,h}$ sobre uma função f, pode ser entendida como a derivada da parábola interpolada a partir dos pontos $(x-h, f(x-h))$, $(x, f(x))$ e $(x+h, f(x+h))$. Verifique esse fato utilizando a interpolação de Lagrange ou Newton.

Erros de truncamento

Vamos analisar os erros de truncamento cometidos quando calculamos numericamente a derivada de uma função através das operações de diferença finita.

A diferença entre $(D_{+,h}f)(x)$ e $f'(x)$ é dada por

$$(D_{+,h}f)(x) - f'(x) = \frac{f(x+h) - f(x)}{h} - f'(x).$$

Através da expansão em série de Taylor em torno de $h = 0$ para $f(x+h)$, notamos que a diferença

[1] A notação $\phi|_x$ para a função ϕ deve ser entendida simplesmente como o valor da função quando seu argumento assume o valor x, i. e., $\phi|_x \equiv \phi(x)$.

assume a forma

$$\begin{aligned}(D_{+,h}f)(x) - f'(x) &= \frac{f(x) + hf'(x) + \frac{h^2}{2}f''(x) + O(h^3) - f(x)}{h} - f'(x) \\ &= \frac{h}{2}f''(x) + O(h^2) = O(h).\end{aligned}$$

De modo análogo, a diferença entre a operação $D_{-,h}$ e a derivada $f'(x)$ também é $O(h)$. Porém, a diferença entre a operação $D_{0,h}f$ e $f'(x)$ é $O(h^2)$:

$$\begin{aligned}(D_{0,h}f)(x) - f'(x) &= \frac{f(x+h) - f(x-h)}{2h} - f'(x) \\ &= \frac{f(x) + hf'(x) + \frac{h^2}{2}f''(x) - f(x) + hf'(x) - \frac{h^2}{2}f''(x) + O(h^3)}{2h} - f'(x) \\ &= O(h^2).\end{aligned}$$

Exemplo 20: Vamos estudar a derivação numérica da função exponencial $f(x) = e^x$, em particular $f(1) = e = 2{,}718281\ldots$ De acordo com a definição dos operadores de diferença finita podemos montar a seguinte tabela:

h	$g_{+,h}(1)$	$g_{-,h}(1)$	$g_{0,h}(1)$	$g_{+,h}(1) - e$	$g_{-,h}(1) - e$	$g_{0,h}(1) - e$
0,4	3,3423	2,2404	2,791352	0,624	-0,478	0,0731
0,2	3,0092	2,4637	2,736440	0,291	-0,254	0,0182
0,1	2,8588	2,5865	2,722815	0,141	-0,131	0,00453
0,05	2,7874	2,6514	2,719414	0,0691	-0,0669	0,00113

onde $g_{+,h}(x) = (D_{+,h}f)(x)$ e a mesma notação é utilizada nas demais aproximações. Os valores da tabela permitem verificar o comportamento do erro de truncamento cometido em cada uma das operações de diferença finita. Enquanto que nas duas primeiras operações, $D_{+,h}$ e $D_{-,h}$ o erro decresce a uma razão de aproximadamente $\frac{1}{2}$ (a razão entre os espaçamentos decresce nessa mesma razão o que está de acordo com a previsão $O(h)$), no caso da operação $D_{0,h}$, o erro decresce a uma razão de aproximadamente $\frac{1}{4}$, o que é consistente com a previsão $O(h^2)$.

Erros de arredondamento

Os erros de truncamento não são os únicos fatores importantes na determinação da estimativa numérica da operação de diferenciação quando essa operação é realizada por máquinas. Nesses casos devemos levar em conta que os números não podem ser armazenados com precisão indefinida, eles são armazenados como um elemento de um sistema de ponto flutuante e todas as operações aritméticas realizadas nesse elemento estão sujeitas a erros de arredondamento.

Vamos tomar como exemplo a operação de diferença finita $D_{+,h}$:

$$(D_{+,h}f)(x) = \frac{f(x+h) - f(x)}{h} = \frac{f_1 - f_0}{h},$$

onde, por economia de notação, representamos $f(x+h) = f_1$ e $f(x) = f_0$. Se a operação for realizada em uma máquina, tipicamente, os valores f_1 e f_0 serão representados por pontos flutuantes $\hat{f}_1$ e $\hat{f}_0$ respectivamente, que correspondem ao resultado das operações realizadas utilizando a aritmética de máquina.

Internamente a função $(D_{+,h}f)(x)$ é representada pelo resultado das operações em ponto flutuante[2] $\left(\hat{f}_1 \ominus \hat{f}_0\right) \oslash h$, cuja diferença em relação ao valor exato $\dfrac{\hat{f}_1 - \hat{f}_0}{h}$ pode ser expressa através da função $\varepsilon(x,h)$:

$$\left(\hat{f}_1 \ominus \hat{f}_0\right) \oslash h = \frac{\hat{f}_1 - \hat{f}_0}{h}\left(1 + \varepsilon(x,h)\right).$$

De modo semelhante, a diferença em valor absoluto entre $\hat{f}_i$ e f_i para uma escolha de x e h é representada por uma função não negativa $\delta(x,h)$. Se h for suficientemente pequeno, teremos então

$$|\hat{f}_1 - f_1| = \delta(x,h) \leq \delta \qquad \text{e} \qquad |\hat{f}_0 - f_0| = \delta(x,0) \leq \delta.$$

A diferença entre o valor da derivada de f em x e o ponto flutuante $\left(\hat{f}_1 \ominus \hat{f}_0\right) \oslash h$ é dada em valor absoluto por

$$\begin{aligned}
\left|f'(x) - \left(\hat{f}_1 \ominus \hat{f}_0\right) \oslash h\right| &= \left|f'(x) - \frac{\hat{f}_1 - \hat{f}_0}{h}\left(1 + \varepsilon(x,h)\right)\right| \\
&= \left|f'(x) - \left(\frac{\hat{f}_1 - \hat{f}_0}{h} + \frac{f_1 - f_1}{h} + \frac{f_0 - f_0}{h}\right)\left(1 + \varepsilon(x,h)\right)\right| \\
&= \left|f'(x) + \left(-\frac{f_1 - f_0}{h} - \frac{\hat{f}_1 - f_1}{h} + \frac{\hat{f}_0 - f_0}{h}\right)\left(1 + \varepsilon(x,h)\right)\right| \\
&\leq \left|f'(x) - \frac{f_1 - f_0}{h}\right| + \left(\left|\frac{\hat{f}_1 - f_1}{h}\right| + \left|\frac{\hat{f}_0 - f_0}{h}\right|\right)|1 + \varepsilon(x,h)| + \\
&+ \left|\frac{f_1 - f_0}{h}\varepsilon(x,h)\right| \\
&= \left|f'(x) - \frac{f_1 - f_0}{h}\right| + \frac{\left(\delta(x,h) + \delta(x,0)\right)|1 + \varepsilon(x,h)|}{h} + \\
&+ \left|\frac{f_1 - f_0}{h}\varepsilon(x,h)\right| \\
&\leq \left(c_1 h + \frac{2\delta}{h}\right)(1 + \varepsilon) + \left|f'(x)\right|\varepsilon \qquad (4.0.1)
\end{aligned}$$

onde $c_1 = \dfrac{1}{2}\max_{x \leq y \leq x+h} f''(y)$ é o erro devido ao truncamento (que independe de h).

Podemos notar pela estimativa (4.0.1) que o erro cometido ao calcularmos numericamente a derivada também cresce quando tomamos valores de h muito pequenos. Isto é um reflexo direto das limitações da aritmética de ponto flutuante utilizadas pela máquina. Portanto, ao utilizar operações de diferença finita em uma máquina para calcular numericamente a derivada de uma função,

[2] Por simplicidade, assumimos que h é idêntico a sua representação em ponto flutuante.

devemos analisar cuidadosamente a escolha de um espaçamento h ótimo. Sempre devemos tomar esse cuidado em qualquer operação de diferença finita.

Na subseção seguinte vamos estudar como desenvolver aproximações mais precisas para a derivada de uma função f.

4.1 Extrapolação de Richardson

No início deste capítulo vimos que podemos estudar os erros cometidos nas operações de diferença finita através da expansão em série de potências de h (o espaçamento entre os pontos). Vamos rever o caso do operador de diferença finita $D_{0,h}$ com um maior número de termos na expansão:

$$\begin{aligned}
\left(D_{0,h}f\right)(x) &:= \frac{f(x+h)-f(x-h)}{2h} \\
&= \frac{1}{2h}\left(f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{3!}f^{(3)}(x)+O(h^4)\right) \\
&\quad -\frac{1}{2h}\left(f(x)-hf'(x)+\frac{h^2}{2}f''(x)-\frac{h^3}{3!}f^{(3)}(x)+O(h^4)\right) \\
&= f'(x)+c_2h^2+O(h^4), \qquad (4.1.1)
\end{aligned}$$

onde $c_2 = \dfrac{f^{(3)}(x)}{3!}$. Portanto, o erro de truncamento até a segunda ordem em h é c_2h^2, onde, naturalmente a constante c_2 não depende de h. Dessa forma, a operação de diferença finita com espaçamento $2h$, $\left(D_{0,2h}f\right)(x)$, é tal que

$$\begin{aligned}
\left(D_{0,2h}f\right)(x) &:= \frac{f(x+2h)-f(x-2h)}{2(2h)} \\
&= f'(x)+c_2(2h)^2+O(h^4). \qquad (4.1.2)
\end{aligned}$$

E portanto a diferença entre a ação dessas duas operações aplicadas a uma função f pode ser descrita na segunda ordem em h como

$$\left(D_{0,2h}f\right)(x)-\left(D_{0,h}f\right)(x)=3c_2h^2+O(h^4).$$

Ou seja, é possível descrever o termo c_2h^2 através das duas operações de diferença finita mais termos de ordem h^4:

$$c_2h^2=\frac{\left(D_{0,2h}f\right)(x)-\left(D_{0,h}f\right)(x)}{3}+O(h^4). \qquad (4.1.3)$$

A substituição do termo (4.1.3) em qualquer das duas expressões (4.1.1) ou (4.1.2) permite expressar a derivada de f em termos de operações de diferença finita envolvendo cinco pontos :

$x-2h, x-h, x, x+h$ e $x+2h$:

$$\begin{aligned} f'(x) &= (D_{0,h}f)(x) - \frac{(D_{0,2h}f)(x) - (D_{0,h}f)(x)}{3} + O(h^4) \\ &= \frac{4(D_{0,h}f)(x) - (D_{0,2h}f)(x)}{3} + O(h^4) \\ &:= (D_{1,h}f)(x) + O(h^4). \end{aligned}$$

Essa técnica é denominada *extrapolação de Richardson*, através dela é possível construir operações de diferença finita com maior precisão. No exemplo que acabamos de estudar determinamos a nova operação de diferença finita $D_{1,h}$ a partir das operações $D_{0,h}$ e $D_{0,2h}$:

$$\begin{aligned} (D_{1,h}f)(x) &:= \frac{4(D_{0,h}f)(x) - (D_{0,2h}f)(x)}{3} \\ &= \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}. \end{aligned}$$

Como acabamos de verificar, $f'(x) - (D_{1,h}f)(x) = O(h^4)$. Portanto, a diferença anterior pode ser escrita como $f'(x) - (D_{1,h}f)(x) = c_4h^4 + O(h^6)$, onde c_4 é também um termo que independe de h. E assim considerando a operação com espaçamento duplo $(D_{1,2h}f)(x) = f'(x) + c_4(2h)^4 + O(h^6)$ podemos dar prosseguimento a extrapolação e determinar a operação $D_{2,h}$ tal que

$$(D_{2,h}f)(x) := \frac{16(D_{1,h}f)(x) - (D_{1,2h}f)(x)}{15}$$

e $f'(x) - (D_{2,h}f)(x) = O(h^6)$.

Exemplo 21: Voltamos a estudar o exemplo do início do capítulo: a derivação numérica da função exponencial $f(x) = e^x$, em particular $f(1) = e = 2{,}718281\ldots$ De acordo com a definição dos operadores de diferença finita podemos montar a seguinte tabela:

h	$(D_{0,h}f)(1)$	$(D_{1,h}f)(1)$	$(D_{2,h}f)(1)$
0,4	2,791352		
0,2	2,736440	2,718136	
0,1	2,722815	2,718273	2,718282
0,05	2,719414	2,718280	2,718281

pelo fato dos espaçamentos serem múltiplos de 0,05 podemos utilizar a informação sobre as operações de ordem menor para calcular as de maior ordem.

De maneira geral podemos enunciar o processo de construção de uma operação de diferença finita de ordem de truncamento mais alta: Seja a operação de diferença finita F_h que aproxima a n-ésima derivada de uma função suficientemente suave g até a ordem $O(h^p)$, ou seja,

$$(F_hg)(x) = g^{(n)}(x) + ch^p + O(h^r),$$

onde $r > p$. Então, para qualquer $q > 1$, segundo o processo de extrapolação de Richardson,

$$(F_h g)(x) + \frac{1}{q^p - 1}\left((F_h g)(x) - (F_{qh} g)(x)\right) = g^{(n)}(x) + O(h^r).$$

O que permite definir a nova operação $\tilde{F}_h$:

$$\left(\tilde{F}_h g\right)(x) := \frac{q^p (F_h g)(x) - (F_{qh} g)(x)}{q^p - 1}.$$

As operações de diferença finita para as segundas derivadas e as derivadas de ordem superior podem ser obtidas a a partir da combinação das operações de diferença finita para primeira derivada, portanto, as operações $D_{2,h}$, $D_{+,h}D_{+,h}$, $D_{-,h}D_{-,h}$, $D_{0,h}D_{+,h}$, $D_{1,h}D_{-,h}$, etc., fornecem aproximações para a segunda derivada com diferentes precisões. Porém a extrapolação de Richardson pode ser utilizada para aumentá-la se for necessário. Por exemplo, a ação da operação $D_{+,h}D_{-,h}$ sobre uma função f é dada por

$$(D_{+,h}D_{-,h}f) = (D_{+,h}g),$$

onde $g(x) = (D_{-,h}g)(x)$, ou seja,

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

e portanto

$$\begin{aligned}
(D_{+,h}g)(x) &= \frac{1}{h}(D_{+,h}f)(x) - \frac{1}{h}(D_{+,h}f(\cdot - h))(x) \\
&= \frac{1}{h}\left(\frac{f(x+h) - f(x)}{h}\right) - \frac{1}{h}\left(\frac{f(x) - f(x-h)}{h}\right) \\
&= \frac{f(x+h) - 2f(x) + f(x-h)}{h^2},
\end{aligned}$$

onde o termo $f(\cdot - x)$ é uma abreviação para a nova função $q(x) = f(x-h)$. Realizando a expansão em série de potências para h podemos verificar que

$$(D_{+,h}D_{-,h}f)(x) = f''(x) + O(h^2).$$

Esse exemplo que acabamos de estudar é apenas uma possibilidade, como acabamos de afirmar, existem diversas outras possibilidades.

Exemplos

Solução estacionária para a equação do calor

4.2 Exercícios

1) Seja o operador diferença finita $D_{+,h}$, definido como

$$\left(D_{+,h}f\right)(x) = \frac{f(x+h) - f(x)}{h}.$$

Sabemos que $\left(D_{+,h}f\right)(x) - f'(x) = O(h)$. Utilize a extrapolação de Richardson para encontrar a expressão do operador $D_{+2,h}$.

2) Dada uma função f tal que

x	0,1	0,2	0,3	0,4	0,5
$f(x)$	0,23	0,33	0,29	0,12	0,17

Tabela 4.1: valores de f

determine estimativas com quatro dígitos para a derivada de f nos pontos $x = 0{,}3$ e $x = 0{,}1$ a partir de todos os valores da tabela.

3) Determine a ordem do erro de truncamento ao aproximar a operação de derivação de segunda ordem pela operação de diferença finita $D_{+,h}D_{+,h}$.

4) Determine a ordem do erro de truncamento ao aproximar a operação de derivação de segunda ordem pela operação de diferença finita $D_{+,h}D_{1,h}$.

5) Encontre a extrapolação de Richardson calculada com 5 pontos ($x-2h$,$x-h$, x, $x+h$ e $x+2h$) para a segunda derivada da função f.

6) Desenvolva uma expressão para a operação finita que aproxima a segunda derivada de uma função f no ponto x a partir de $f(x)$, $f(x+h)$ e $f(x+2h)$.

7) Construa todas as possíveis operações de diferença finita associadas às derivadas de f no ponto x a partir de combinações lineares dos termos $f(x)$, $f(x+h)$ e $f(x+2h)$.

8) Um cilindro com paredes laterais isoladas termicamente possui uma das bases em contato com um material de propriedades térmicas diferentes. Três termômetros inseridos no cilindro fornecem uma leitura da sua temperatura em pontos que distam 2cm , 4cm e 6cm da base. Dado que:

- as temperaturas mantêm-se constantes ao longo do tempo,
- a condutividade térmica do cilindro, k, é de $150\frac{\mathrm{W}}{\mathrm{m\ K}}$,
- a temperatura medida nos pontos vale respectivamente 350K, 358K e 373K,

a partir da Lei de Fourier

$$\vec{q} = -k\nabla T$$

e de uma operação de diferença finita, determine uma estimativa com quatro dígitos para o módulo do fluxo térmico pela base do cilindro.

9) Utilize a extrapolação de Richardson para construir um aproximação com erro de truncamento $O\left(h^4\right)$ a partir da operação de diferença finita

$$D_h f\left(x\right) = \frac{-f(x-2h) + 2f(x-h) - 2f(x+h) + f(x+2h)}{2h^3} = f'''(x) + O\left(h^2\right).$$

10) Uma função suave $f(x)$ é conhecida em x, $x+2h$ e $x-h$ para um dado $h > 0$. Determine uma aproximação para $f'(x)$ a partir do valor de f nesses três pontos.

11) O operador de diferença finita Δ_h aproxima o seguinte operador diferencial

$$(\Delta_h f)(x) := \frac{1}{30h}\left(-8f(x-2h) + 5f(x+h) + 3f(x+3h)\right) = f'(x) + O\left(h^2\right).$$

Utilize a extrapolação de Richardson para construir a forma explícita de um operador com erro de truncamento de ordem superior em h.

12) A partir dos valores de uma função y, conhecida nos pontos uniformemente distribuídos $(\ldots, x-2h, x-h, x)$, construa uma operação de diferença finita que aproxime o valor de

$$y'(x) - 2y''(x)$$

com erro de truncamento $O\left(h^2\right)$.

13) Uma esfera uniforme de raio R possui a superfície externa em contato com um material de propriedades térmicas diferentes. Em um determinado instante, a temperatura depende apenas da distância r ao centro e é conhecida nas distâncias R, $\frac{5}{6}R$ e $\frac{4}{6}R$. Utilize essas informações para determinar uma aproximação para $\frac{dT}{dr}(R)$ a partir de uma operação de diferença finita.

5 Interpolação

Neste capítulo estudaremos métodos que permitem encontrar um valor aproximado para uma função f calculada em um ponto x do intervalo I, através do conhecimento de uma coleção de pares ordenados (pontos) $\{(x_i, f(x_i))\}_{i=1}^{N}$tais que $x_i \in I$. Seja g uma função que aproxima f no intervalo I. Então, para o conjunto de pontos x_i, $i = 1, \ldots, N$

$$g(x_i) = f(x_i),$$

dizemos que g *interpola* a função f nos valores $x_1, x_2, \ldots, x_N$. Então podemos utilizar a função g para encontrar uma aproximação para o valor de f no ponto[1] $x \in [x_1, x_n]$, esse procedimento é denominado *interpolação*. Se x estiver fora do intervalo $[x_1, x_n]$ e ainda assim utilizarmos a função g para encontrar o valor aproximado de f nesse ponto, o procedimento é denominado *extrapolação*.

Exemplo 22: Vamos determinar uma função interpolante para o conjunto de pontos $\{(-0{,}5; -5{,}0); (0{,}5; 0{,}81); (1{,}0; 0{,}7); (1{,}5; 0{,}55)\}$ na forma $g(x) = a_1 e^{a_2 x} + a_3 e^{a_4 x}$. Determinar o valor dos coeficientes a_1, a_2, a_3 e a_4 significa determinar a interpolação.

Por definição, se g interpola o conjunto de pontos (entendido como $\{(x_i, f(x_i))\}_{i=1}^{4}$) então os coeficientes devem satisfazer as quatro equações $g(x_1) = f(x_1)$, ..., $g(x_4) = f(x_4)$, ou seja, devem ser solução do seguinte sistema de equações não lineares:

$$\begin{cases} a_1 e^{-a_2 0{,}5} + a_3 e^{-a_4 0{,}5} &= -5{,}0 \\ a_1 e^{a_2 0{,}5} + a_3 e^{a_4 0{,}5} &= 0{,}81 \\ a_1 e^{a_2} + a_3 e^{a_4} &= 0{,}7 \\ a_1 e^{a_2 1{,}5} + a_3 e^{a_4 1{,}5} &= 0{,}55 \end{cases}.$$

Esse sistema possui solução numérica dada por

$$\begin{aligned} a_1 &\approx 1{,}20334 \\ a_2 &\approx -0{,}519387 \\ a_3 &\approx -0{,}880292 \\ a_4 &\approx -4{,}01704 \end{aligned}$$

[1] Supondo que os pontos $x_1, x_2, \ldots, x_n$ estão ordenados

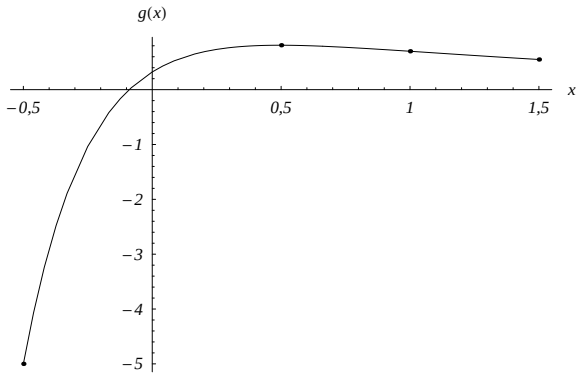


Figura 5.0.1: Função interpolante $g(x)$ e os quatro pontos de interpolação.

O sucesso em conseguir determinar a interpolação de um conjunto de pontos depende da escolha de função interpolante. No exemplo anterior a interpolação foi possível pois o "comportamento" dos pontos é compatível com a escolha realizada para a função interpolante. Essa "compatibilidade" se manifesta na existência de solução para o sistema de equações associado à interpolação. Se fosse escolhida uma função com comportamento muito distinto do manifestado pelos pontos, o sistema resultante poderia não possuir solução.

A escolha de polinômios como funções interpolantes é natural pelos seguintes motivos: é possível aproximar uma grande variedade de funções, os polinômios são de fácil manipulação matemática (principalmente derivação e integração) e o Teorema da Aproximação de Weierstrass

Teorema 5.0.1 (**Weierstrass**)
Seja f uma função contínua definida no intervalo fechado limitado $[a, b]$ e seja δ um número positivo. Então existe um polinômio p, tal que para todo $x \in [a, b]$,

$$|f(x) - p(x)| < \delta.$$

No entanto, da mesma forma que o teorema de Weierstrass garante uma representação de f por um polinômio p tão "próximo" quanto queiramos, ele nada diz sobre o grau de p. Em algumas situações, o problema de encontrar p que desempenhe esse papel pode ser extraordinariamente difícil do ponto de vista numérico.

Antes de discutirmos o procedimento de interpolação por polinômios, vale a pena mencionar um algoritmo útil no cálculo do valor de p em um ponto x. Trata-se do algoritmo de Horner.

Algoritmo de Horner

Batizado com o nome do matemático inglês Willian George Horner mas já conhecido por Isaac Newton em 1669 e mesmo pelo matemático chinês Qin Jiunshao no séc. XIII, o algoritmo fornece uma maneira otimizada de calcular $p(x) = a_m x^m + a_{m-1} x^{m-1} + \ldots + a_1 x + a_0$ através de m multiplicações e m adições.

Partindo do polinômio reescrito na forma concatenada

$$p(x) = ((\ldots((a_m x + a_{m-1})x + a_{m-2})x + \ldots + a_2)x + a_1)x + a_0$$

introduzimos as constantes b_j definidas recursivamente por

$$b_1 := a_m \qquad \text{e} \qquad b_{j+1} := b_j x + a_{m-j}, \text{ a para } j = 1, 2, \ldots, m. \tag{5.0.1}$$

O valor que p assume em um dado x corresponde à constante b_{m+1}, ou seja, $p(x) = b_{m+1}$, cujo valor é o resultado das m multiplicações e m adições presentes em (5.0.1).

Como exemplo, vamos considerar o polinômio $p(x) = 3x^3 + 8x^2 - x + 1$. Neste caso, dado um valor x

$$\begin{aligned} b_1 &= 3 \\ b_2 &= b_1 x + a_2 &&= 3x + 8 \\ b_3 &= b_2 x + a_1 &&= (3x + 8)x - 1 \end{aligned}$$

e finalmente

$$p(x) = b_4 = b_3 x + a_0 = ((3x + 8)x - 1) + 1.$$

5.1 Interpolação polinomial

Seja $f_i, i = 1, 2, \ldots, n$, o valor da função f calculada nos n pontos de interpolação x_i. Encontrar o polinômio de grau m que interpola f nesses pontos consiste em resolver o sistema de equações lineares $f_i \equiv f(x_i) = p(x_i)$, ou seja o sistema

$$\left\{ \begin{array}{ccccccccccc} a_m x_1^m & + & a_{m-1} x_1^{m-1} & + & \ldots & + & a_1 x_1 & + & a_0 & = & f_1 \\ a_m x_2^m & + & a_{m-1} x_2^{m-1} & + & \ldots & + & a_1 x_2 & + & a_0 & = & f_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_m x_n^m & + & a_{m-1} x_n^{m-1} & + & & + & a_1 x_n & + & a_0 & = & f_n \end{array} \right. \tag{5.1.1}$$

As $m+1$ incógnitas são os coeficientes do polinômio, $a_0, a_1, \ldots, a_m$ e o sistema possui n equações. Portanto, tipicamente, o sistema não possui solução se $m + 1 < n$, possui infinitas soluções se $m + 1 > n$ e será unicamente determinado se $m + 1 = n$.

5.1.1 Interpolação pelos polinômios de Lagrange

Como veremos adiante, resolver o sistema (5.1.1) não é a maneira mais simples ou menos sujeita a erros de arredondamento quando desejamos determinar o polinômio interpolante. O seguinte teorema garante a unicidade do polinômio interpolante, o que nos permite buscar maneiras alternativas de construí-lo. Por ser único, o resultado será independente da construção.

Teorema 5.1.1 (unicidade do polinômio interpolante)
Sejam $x_1, \ldots, x_n$, pontos distintos. Para um conjunto arbitrário de valores $f_1, \ldots, f_n$ existe um e somente um polinômio p de grau menor ou igual a $n - 1$ tal que

$$p(x_i) = f_i,$$

para $i = 1, 2, \ldots, n$.

Demonstração: No caso em que temos n pontos distintos e procuramos um polinômio de grau menor ou igual a $n - 1$, a matriz quadrada dos coeficientes do sistema de equações

lineares (5.1.1) assume a forma da seguinte matriz de Vandermonde,

$$\begin{pmatrix} x_1^{n-1} & x_1^{n-2} & \cdots & x_1^2 & x_1 & 1 \\ x_2^{n-1} & x_2^{n-2} & \cdots & x_2^2 & x_2 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ x_n^{n-1} & x_n^{n-2} & \cdots & x_n^2 & x_n & 1 \end{pmatrix}.$$

Por hipótese os x_i são distintos, portanto o determinante da matriz, dado por

$$\prod_{1 \leq i < j \leq n} (x_j - x_i)$$

é não nulo, consequentemente, a solução do sistema é única e o polinômio também. ■

Vamos supor que para cada $1 \leq j \leq n$ exista um polinômio de grau $n-1$, $l_j(x)$ tal que para cada $1 \leq k \leq n$, o valor de l_j no ponto de interpolação x_k é tal que

$$l_j(x_k) = \delta_{j,k},$$

onde $\delta_{j,k}$ é o delta de Kronecker[2]. Nesse caso, os polinômios l_j permitem reescrever o polinômio interpolante $p(x)$:

$$p(x) = f_1 l_1(x) + f_2 l_2(x) + \ldots + f_n l_n(x) = \sum_{j=1}^{n} f_j l_j(x),$$

podemos trivialmente verificar que $p(x_k) = \sum_{j=1}^{n} f_j l_j(x_k) = \sum_{j=1}^{n} f_j \delta_{j,k} = f_k$. Portanto se formos capazes de construir os polinômios l_j a interpolação estará determinada. Vamos então construí-los a partir das seguintes considerações.

Segundo a sua definição $l_j(x_k) = 0$ para todo x_k tal que $k \neq j$, então os pontos x_k são raízes de l_j, se $j \neq k$ e portanto, a menos de uma constante multiplicativa, C_j, o polinômio l_j é determinado pelo produtório

$$\begin{aligned} l_j(x) &= C_j(x - x_1)(x - x_2)\ldots(x - x_{j-1})(x - x_{j+1})\ldots(x - x_n) \\ &= C_j \prod_{\substack{i=1 \\ i \neq j}}^{n} (x - x_i). \end{aligned}$$

Por fim, a constante C_j pode ser determinada através da propriedade $l_j(x_j) = 1$:

$$l_j(x_j) = 1 \quad \Rightarrow \quad C_j \prod_{\substack{i=1 \\ i \neq j}}^{n} (x_j - x_i) = 1,$$

[2] O delta de Kronecker é definido pela expressão

$$\delta_{j,k} = \begin{cases} 0 & , \quad j \neq k \\ 1 & , \quad j = k \end{cases},$$

onde j e k são dois números inteiros.

ou seja

$$C_j = \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{1}{(x_j - x_i)}.$$

Dessa forma, os polinômios $l_j(x)$, denominados polinômios de Lagrange são determinados a partir do seguinte produtório

$$l_j(x) = \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}$$

e a interpolação de Lagrange

$$p(x) = \sum_{j=1}^{n} f_j l_j(x).$$

Exemplo 23: Seja a função $f(x) = \text{sen}(x)$ a partir da qual construímos a interpolação nos três pontos $x_1 = 0$, $x_2 = 1$ e $x_3 = 2$. Será então um polinômio de segundo grau. Os pontos de interpolação são dados por

j	x_j	$f_j = \text{sen}(x_j)$
1	0	0
2	1	sen(1)
3	2	sen(2)

os polinômios de Lagrange são então dados por

$$l_1(x) = \frac{(x-1)(x-2)}{(0-1)(0-2)} = \frac{x^2 - 3x + 2}{2},$$

$$l_2(x) = \frac{(x-0)(x-2)}{(1-0)(1-2)} = -x^2 + 2x$$

e

$$l_3(x) = \frac{(x-0)(x-1)}{(2-0)(2-1)} = \frac{x^2 - x}{2}.$$

A interpolação é dada por

$$p(x) = \left(\frac{\text{sen}(2)}{2} - \text{sen}(1)\right) x^2 + \left(2\text{sen}(1) - \frac{\text{sen}(2)}{2}\right) x$$

5.1.2 Interpolação de Newton

De acordo com o teorema da unicidade do polinômio interpolante, toda interpolação de n pontos por um polinômio de grau $n-1$ é única e pode ser obtida pelo método de Lagrange. No entanto, existem outras maneiras de construir o polinômio $p(x)$ que podem ser mais convenientes. Uma dessas maneiras é a interpolação de Newton, que permite a inserção de pontos adicionais de maneira simples e menos suscetível à deterioração por erros de arredondamento.

O método consiste em determinar o polinômio

$$p(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)(x - x_2) + \ldots + a_{n-1}(x - x_1)\ldots(x - x_{n-1}).$$

Por construção, o valor de p calculado em $x = x_1$ é

$$p(x_1) = a_0.$$

Além disso, como $p(x)$ é o polinômio interpolante, $p(x_1) = f_1$, portanto,

$$a_0 = f_1.$$

Da mesma forma,

$$\begin{aligned} p(x_2) &= a_0 + a_1(x_2 - x_1) = f_2 \\ &= f_1 + a_1(x_2 - x_1) = f_2, \end{aligned}$$

ou seja,

$$a_1 = \frac{f_2 - a_0}{x_2 - x_1}$$

e assim por diante, os coeficientes são determinados recursivamente e o k-ésimo coeficiente é determinado em função dos pontos de interpolação e dos coeficientes anteriores pela expressão

$$a_k = \frac{f_{k+1} - a_0 - \sum_{j=1}^{k-1} a_j (x_{k+1} - x_1) \dots (x_{k+1} - x_j)}{\prod_{j=1}^{k} (x_{k+1} - x_j)}. \tag{5.1.2}$$

A fórmula de recorrência (5.1.2) pode ser convenientemente descrita através da notação de ***diferenças divididas***. Seja a função $f[x_k, x_{k+1}, \dots, x_{l+1}]$ definida pela relação de recorrência

$$f[x_k, x_{k+1}, \dots, x_l, x_{l+1}] = \frac{f[x_{k+1}, x_{k+2}, \dots, x_{l+1}] - f[x_k, x_{k+1}, \dots, x_l]}{x_{l+1} - x_k}$$

e

$$f[x_k] = f_k := f(x_k).$$

Assim, podemos verificar que

$$f[x_k, x_{k+1}] = \frac{f[x_{k+1}] - f[x_k]}{x_{k+1} - x_k}$$

e

$$f[x_k, x_{k+1}, x_{k+2}] = \frac{f[x_{k+1}, x_{k+2}] - f[x_k, x_{k+1}]}{x_{k+2} - x_k}.$$

Nessa notação, os coeficientes do polinômio são dados por

$$\begin{aligned} a_0 &= f[x_1], \\ a_1 &= f[x_1, x_2], \\ a_2 &= f[x_1, x_2, x_3], \\ &\vdots \quad \vdots \quad \vdots \\ a_{n-1} &= f[x_1, x_2, \dots, x_n]. \end{aligned}$$

Diagramaticamente, os coeficientes são calculados a partir da sequência de diferenças divididas calculadas recursivamente:

$$
\begin{array}{ccccccccccc}
x_1 & \to & f[x_1] & \to & f[x_1,x_2] & \to & f[x_1,x_2,x_3] & \dots & f[x_1,\dots,x_{n-1}] & \to & f[x_1,\dots,x_n] \\
 & & & \nearrow & & \nearrow & & & & \nearrow & \\
x_2 & \to & f[x_2] & \to & f[x_2,x_3] & \to & f[x_2,x_3,x_4] & \dots & f[x_2,\dots,x_n] & & \\
 & & & \nearrow & & \nearrow & & & & & \\
x_3 & \to & f[x_3] & \to & f[x_3,x_4] & \to & f[x_3,x_4,x_5] & \dots & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & & \\
x_{n-2} & \to & f[x_{n-2}] & \to & f[x_{n-2},x_{n-1}] & \to & f[x_{n-2},x_{n-1},x_n] & \cdots & & & \\
 & & & \nearrow & & \nearrow & & & & & \\
x_{n-1} & \to & f[x_{n-1}] & \to & f[x_{n-1},x_n] & & & & & & \\
 & & & \nearrow & & & & & & & \\
x_n & \to & f[x_n] & & & & & & & &
\end{array}
$$

Exemplo 24: Vamos realizar a interpolação da função $\text{sen}(x)$ no intervalo $x \in [0,2]$ através de um polinômio de segundo grau nos pontos $x_1 = 0$, $x_2 = 1$ e $x_3 = 2$. Neste caso,

j	x_j	$f_j = \text{sen}(x_j)$
1	0	0
2	1	$\text{sen}(1)$
3	2	$\text{sen}(2)$

e $f[x_1] = 0$, $f[x_2] = \text{sen}(1)$ e $f[x_3] = \text{sen}(2)$. As próximas diferenças divididas são dadas por

$$f[x_1,x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1} = \frac{\text{sen}(1)-0}{1-0}$$

e

$$f[x_2,x_3] = \frac{f[x_3]-f[x_2]}{x_3-x_2} = \frac{\text{sen}(2)-\text{sen}(1)}{2-1}.$$

Finalmente,

$$f[x_1,x_2,x_3] = \frac{f[x_2,x_3]-f[x_1,x_2]}{x_3-x_1} = \frac{\text{sen}(2)-\text{sen}(1)-\text{sen}(1)}{2-0}.$$

Portanto, o polinômio interpolante

$$p(x) = f[x_1] + f[x_1,x_2]\,(x-x_1) + f[x_1,x_2,x_3]\,(x-x_1)(x-x_2)$$

é

$$p(x) = \text{sen}(1)\,x + \frac{\text{sen}(2)-2\text{sen}(1)}{2}\,x(x-1)$$

Exercício 5.1.2. *1) Inclua o ponto $x_4 = 1/2$ na interpolação anterior e encontre o polinômio interpolante de terceiro grau.*

2) Encontre o polinômio interpolante de terceiro grau nos mesmos pontos do exemplo anterior (incluindo o ponto $x_4 = 1/2$) para as funções $\cos(x)$, $xsen(x)$ e $e^x - 1$.

Exemplo 25: A solubilidade do clorato de potássio em água $KClO_3$ nas temperaturas de $0°C$,

$10°C$, $20°C$, $30°C$ e $40°C$ é de $3{,}3g$, $5{,}2g$, $7{,}3g$, $10{,}1g$ e $13{,}9g$ por $100g$ de H_2O, respectivamente.

Nosso objetivo é estabelecer uma boa aproximação para o valor da solubilidade a $25°C$. Vamos inicialmente aproximar a solubilidade a partir de três valores de temperatura. Seja portanto, $\{(x_i, f_i)\}_{i=1}^{3}$ dado por $\{(10; 5{,}2), (20; 7{,}3), (30, 10{,}1)\}$. De acordo com o método de Newton, devemos determinar as diferenças divididas $f[x_1]$, $f[x_1, x_2]$, $f[x_1, x_2, x_3]$ e construir o polinômio

$$p(x) = a_0 + a_1(x - 10) + a_2(x - 10)(x - 20),$$

onde $a_0 = f[x_1]$, $a_1 = f[x_1, x_2]$ e $a_2 = f[x_1, x_2, x_3]$.

$$f[x_1] \equiv f_1 = 5{,}2$$

e $a_0 = 5{,}2$.

$$f[x_1, x_2] = \frac{f_2 - f_1}{x_2 - x_1} = \frac{7{,}3 - 5{,}2}{20 - 10} = 0{,}21$$

e $a_1 = 2{,}1 \times 10^{-1}$.

$$f[x_1, x_2, x_3] = \frac{f[x_2 - x_3] - f[x_1, x_2]}{x_3 - x_1} = \frac{f[x_2 - x_3] - 0{,}21}{30 - 10}$$

Para calcular essa quantia será necessário determinar também o valor de $f[x_2, x_3]$.

$$f[x_2, x_3] = \frac{f_3 - f_2}{x_3 - x_2} = \frac{10{,}1 - 7{,}3}{30 - 20} = 0{,}28$$

Assim,

$$[x_1, x_2, x_3] = \frac{0{,}28 - 0{,}21}{30 - 10} = 0{,}0035$$

e $a_2 = 3{,}5 \times 10^{-3}$.

Portanto a interpolação é dada por

$$p(x) = 5{,}2 + 2{,}1 \times 10^{-1}(x - 10) + 3.5 \times 10^{-3}(x - 10)(x - 20).$$

E a aproximação é dada por

$$p(25) = 8{,}6125.$$

Como a tabela fornece valores com 2 ou três dígitos significativos, a aproximação deverá conter esse mesmo número de dígitos, ou seja, a solubilidade a $25°C$ é de aproximadamente $8{,}6g$ de $KClO_3$por $100g$ de H_2O. Uma maneira de estabelecer a validade dessa aproximação é incluir mais termos à interpolação e verificar se a nova aproximação coincide. Vamos então incluir o valor da solubilidade a $40°C$ como um quarto dado, ou seja, vamos incluir o ponto $(x_4, f_4) = (40; 13{,}9)$. Nesse caso, como (x_1, f_1), (x_2, f_2) e (x_3, f_3) são os mesmos do polinômio anterior, a_0, a_1 e a_2 serão os mesmos no novo polinômio:

$$\tilde{p}(x) = 5{,}2 + 2{,}1 \times 10^{-1}(x - 10) + 3.5 \times 10^{-3}(x - 10)(x - 20) + a_3(x - 10)(x - 20)(x - 30),$$

onde $a_3 = f[x_1, x_2, x_3, x_4]$. Vamos então determinar $f[x_1, x_2, x_3, x_4]$.

$$f[x_1, x_2, x_3, x_4] = \frac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1} = \frac{f[x_2, x_3, x_4] - 0{,}0035}{40 - 10},$$

$$f[x_2, x_3, x_4] = \frac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2} = \frac{f[x_3, x_4] - 0{,}28}{40 - 20}$$

e

$$f[x_3, x_4] = \frac{f_4 - f_3}{x_4 - x_3} = \frac{13{,}9 - 10{,}1}{40 - 30} = 0{,}38.$$

Dessa forma, substituindo os valores

$$f[x_2, x_3, x_4] = \frac{0{,}38 - 0{,}28}{20} = 0{,}005$$

e

$$f[x_1, x_2, x_3, x_4] = \frac{0{,}005 - 0{,}0035}{30} = 5 \times 10^{-5}.$$

O novo polinômio é

$$\tilde{p}(x) = 5{,}2 + 2{,}1 \times 10^{-1}(x-10) + 3{,}5 \times 10^{-3}(x-10)(x-20) + 5 \times 10^{-5}(x-10)(x-20)(x-30)$$

e a aproximação é determinada a partir de

$$\tilde{p}(25) = 8{,}59375\ldots$$

Ou seja, obtemos novamente a aproximação de dois dígitos, $8{,}6g$ de clorato por $100g$ de água.

5.1.3 Erros de truncamento na interpolação por polinômios

Seja f uma função contínua e n vezes diferenciável no intervalo (a, b) que contém os pontos $x_1, x_2, \ldots, x_n$ e seja p o polinômio de grau $n - 1$ que interpola f nesses pontos. Então é possível mostrar[3] que para cada $x \in (a, b)$, existe um $\zeta(x) \in (a, b)$ tal que

$$f(x) - p(x) = \frac{1}{n!} f^{(n)}(\zeta) \prod_{i=1}^{n} (x - x_i). \tag{5.1.3}$$

Poderíamos supor que para uma f contínua e suficientemente suave, a sequência de polinômios interpolantes $\{p_n\}_{n \geq 1}$ convergiria para f conforme aumentássemos o número de pontos de interpolação no intervalo (a, b). No entanto, como o exemplo a seguir ilustra, isto nem sempre ocorre.

[3] A demonstração pode ser encontrada nas referências:
Eldén, L.; Wittmeyer-Koch, L. *Numerical Analysis* (1990),
Claudio, D. M.; Marins, J. M. *Cálculo Numérico Computacional - teoria e prática* 3ªed. (2000).

Fenômeno de Runge

A seguinte função, proposta por Carle D. T. Runge ao estudar o comportamento dos erros na interpolação polinomial,

$$f(x) = \frac{1}{1+25x^2}, \qquad x \in [-1,1]$$

é tal que a sequência de polinômios interpolantes $\{p_n\}_n$ construídos a partir de pontos de interpolação igualmente espaçados não converge[4]para $f(x)$ no intervalo de valores $x \in (-1;-0{,}727)\bigcup(0{,}727;1)$. Na realidade é possível demonstrar que

$$\lim_{n\to+\infty} \max_{-1\leq x\leq 1} |f(x) - p_n(x)| = +\infty.$$

Podemos analisar esse comportamento não regular da interpolação a partir do termo

$$\prod_{i=1}^{n}(x - x_i) \tag{5.1.4}$$

contido na expressão (5.1.3). Esse produtório possui uma flutuação para os valores do argumento próximos à fronteira do intervalo $(-1,1)$ que é progressivamente ampliada conforme aumentamos o número de pontos se os mesmos forem igualmente espaçados. Os gráfico seguintes ajudam a ilustrar o comportamento do produtório (5.1.4).

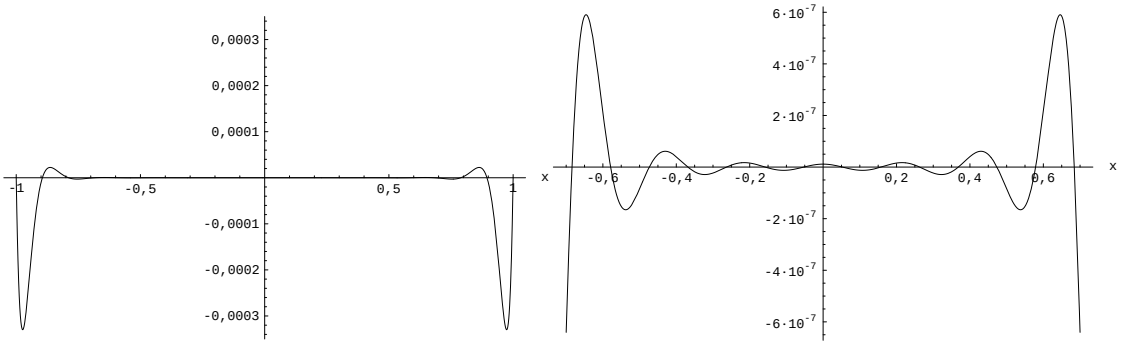


Figura 5.1.1: a) comportamento do produtório (5.1.4) com 20 pontos igualmente espaçados no intervalo $[-1,1]$. b) recorte do mesmo produtório no intervalo $[-0{,}7;0{,}7]$.

A amplitude das oscilações pode ser minimizada através de uma escolha adequada de pontos não uniformemente espaçados. Na realidade, é possível demonstrar que a amplitude de oscilação do termo (5.1.4) é mínima quando os pontos x_i estão espaçados em um intervalo (a,b) segundo a seguinte expressão

$$x_i = \frac{a+b}{2} + \frac{a-b}{2}\cos\left(\frac{2i-1}{2n}\pi\right)$$

para $i = 1,2,\ldots,n$. Esses pontos são denominados *nós de Chebyshev* e consistem em uma transformação afim dos zeros do polinômio de Chebyshev de primeira espécie[5] e grau n no intervalo $(-1,1)$.

[4]A demonstração pode ser encontrada na referência :
Isaacson, E. ; Keller, H. *Analysis of Numerical Methods* (1966).

[5]O polinômio de Chebyshev de primeira espécie e grau n, simbolizado por $T_n(x)$ é uma das duas soluções linearmente

Utilizando os nós de Chebyshev no intervalo $[-1, .1]$ podemos controlar o comportamento dos polinômios interpolantes para a função de Runge e garantir a convergência $p_{n-1}(x) \to f(x)$ quando $n \to +\infty$. De fato, o máximo para o valor absoluto de $\prod_{i=1}^{n}(x - x_i)$ no intervalo $[-1, 1]$ é 2^{1-n} quando x_i são nós de Chebyshev.

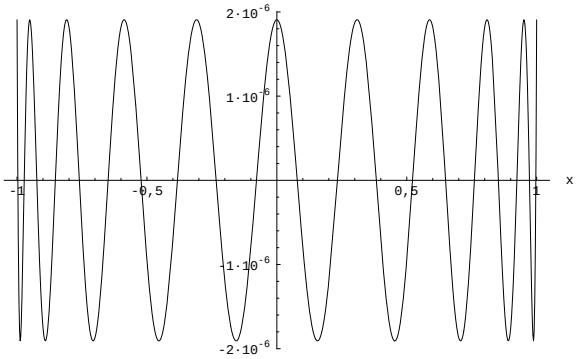


Figura 5.1.2: O produtório (5.1.4) com 20 pontos de Chebyshev

Ainda assim, existem funções contínuas que requerem um número impraticável de pontos para que a interpolação se aproxime da função original. Por exemplo, a função $\sqrt{|x|}$ no intervalo $[-1, 1]$ requer um polinômio de grau maior que 10^6 para que a interpolação seja exata até 10^{-3}.

Em geral, quando utilizamos polinômios de grau maior ou igual a 100, a maior dificuldade é lidar com os erros de arredondamento.

5.2 Interpolação *spline*

Splines são funções formadas por diferentes polinômios de grau menor ou igual a um m, definidos para cada intervalo entre os pontos de interpolação de modo que em cada ponto de interpolação o spline é contínuo,assim como todas as derivadas até ordem $m - 1$.

---

independentes da equação diferencial de Chebyshev,

$$\left(1 - x^2\right) y'' - xy' + \alpha^2 y = 0$$

quando $\alpha = n \in \{0, 1, 2 \ldots\}$. O polinômio pode ser construído a partir das relações de recorrência

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad \text{para } n = 1, 2, \ldots;$$

ou ainda, a partir da fórmula explícita

$$T_n(x) = \cos\left(n \cos^{-1}(x)\right), \qquad x \in [-1, 1].$$

A partir dessa última fórmula, não é difícil demonstrar que os zeros pertencem ao intervalo $(-1, 1)$ e são da forma

$$x_i = \cos\left(\frac{2i - 1}{2n}\pi\right), \qquad i = 1, 2, \ldots, n.$$

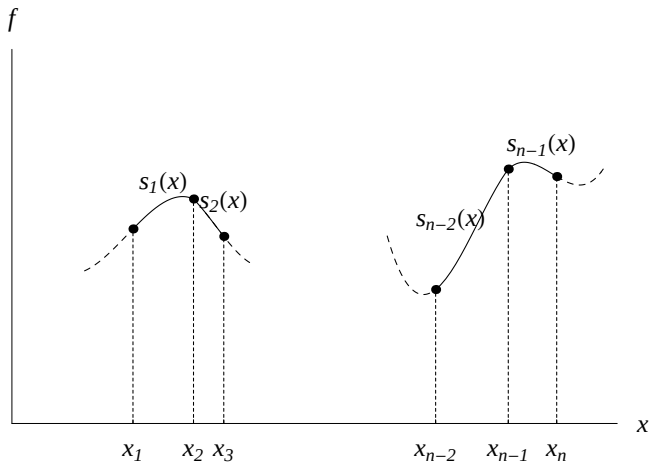


Figura 5.2.1: Interpolação spline

Nas situações em que o número de pontos de interpolação é grande (por exemplo, em aplicações CAD – *computer-aided design*), a inexatidão na aproximação obtida com um polinômio de grau elevado é dominada pelos erros de arredondamento. Ou então quando a função que se quer interpolar possui derivadas de valor numérico elevado em alguma região do intervalo de interpolação, a aproximação é prejudicada em todo o intervalo. Nessas situações, a interpolação por spline pode auxiliar a tarefa de interpolação.

O procedimento de construir splines é análogo qualquer que seja o grau dos polinômios utilizados, como o spline de maior interesse (veremos porque) é aquele formado por polinômios de grau 3, nos concentraremos nesse caso apenas.

5.2.1 Interpolação spline cúbica

Sejam $x_1 < x_2 < \ldots < x_n$ os pontos e interpolação. um spline cúbico é uma função $s(x)$, definida no intervalo $[x_1, x_n]$ com as seguintes propriedades:

1. $s(x)$, $s'(x)$ e $s''(x)$ são funções contínuas no intervalo (x_1, x_n).
2. Em cada subintervalo $[x_i, x_{i+1}]$, $s(x)$ é um polinômio cúbico tal que $s(x_i) = f_i := f(x_i)$ para $i = 1, 2, \ldots, n$.

Portanto, s é composto por $n-1$ polinômios cúbicos, cada polinômio é determinado por 4 coeficientes (a_i,b_i, c_i e d_i) o que dá um total de $4n-4$ coeficientes a determinar, ou seja $4n-4$ incógnitas. Cada polinômio deve satisfazer a condição de continuidade nos pontos de interpolação além, é claro, de interpolar o ponto x_i, ou seja,

$$s_i(x_i) = f_i \qquad \text{(interpolação)},$$

para $i = 1, 2, \ldots, n-1$ e

$$s_{n-1}(x_n) = f_n.$$

A continuidade é satisfeita se

$$s_i(x_{i+1}) = f_{i+1} \quad \text{(continuidade de } s\text{)},$$

para $i = 1, 2, \ldots, n-2$.. As condições acima implicam $2(n-1)$ equações. Faltam ainda as continuidades de $s'(x)$ e $s''(x)$:

$$s_i'(x_{i+1}) = s_{i+1}'(x_{i+1}) \quad \text{(continuidade de } s' \text{)},$$

$$s_i''(x_{i+1}) = s_{i+1}''(x_{i+1}) \quad \text{(continuidade de } s'' \text{)},$$

para $i = 1, 2, \ldots, n-2$. Cada condição equivale a $n-2$ equações. Portanto temos até agora um total de $4n-6$ equações. Restam duas equações para que seu número seja igual ao número de incógnitas. Essas duas últimas equações relacionam-se com as condições de fronteira do spline. Com relação ao comportamento de $s(x)$ no extremo do intervalo, há duas possibilidades a se considerar:

i) spline natural,

$$s_1''(x_1) = 0$$

$$s_{n-1}''(x_n) = 0$$

possui esse nome por ser a condição equivalente à aproximação por réguas elásticas (uso mais tradicional do spline).

ii) spline com mesmas condições de f na extremidade,

$$s_1'(x_1) = f'(x_1)$$

$$s_{n-1}'(x_n) = f'(x_n)$$

essa escolha pressupõe que a informação sobre o valor da derivada de f nos extremos do intervalo seja conhecida. A aproximação obtida com essa escolha possui uma maior exatidão do que a obtida com o spline natural.

Nos próximos parágrafos montaremos o sistema de equações lineares para determinarmos $4n-4$ os coeficientes a_i,b_i, c_i e d_i dos $n-1$ polinômios que compõe o spline:

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3. \tag{5.2.1}$$

Por ser uma interpolação, a cada x_i, temos que $s(x_i) = f_i$, ou seja, $s_i(x_i) = f_i$. Portanto, em vista da equação (5.2.1) a interpolação implica

$$f_i = s_i(x_i) = a_i$$

para $i = 1, 2, \ldots, n-1$. O que determina o valor dos coeficientes a_i.

A continuidade do spline $s(x)$ nos pontos de interpolação implica a equação $s_i(x_{i+1}) = s_{i+1}(x_{i+1})$

para $i = 1, 2, \ldots, n-2$, ou seja,

$$a_i + b_i(x_{i+1} - x_i) + c_i(x_{i+1} - x_i)^2 + d_i(x_{i+1} - x_i)^3 = a_{i+1},$$

$$f_i + b_i(x_{i+1} - x_i) + c_i(x_{i+1} - x_i)^2 + d_i(x_{i+1} - x_i)^3 = f_{i+1}. \tag{5.2.2}$$

Para aliviar a notação, vamos introduzir a notação $h_i = (x_{i+1} - x_i)$. Dessa forma, a equação anterior (5.2.2) pode ser reescrita como

$$f_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = f_{i+1} \tag{5.2.3}$$

A continuidade na primeira e na segunda derivadas implicam

$$b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} \tag{5.2.4}$$

e

$$c_i + 3d_i h_i = c_{i+1} \tag{5.2.5}$$

para $i = 1, 2, \ldots, n-2$.

Isolando d_i na equação (5.2.5) e substituindo em (5.2.3) e (5.2.4) encontramos respectivamente

$$f_{i+1} = f_i + b_i h_i + \frac{h_i^2}{3}(2c_i + c_{i+1}) \tag{5.2.6}$$

e

$$b_{i+1} = b_i + h_i(c_i + c_{i+1}) \tag{5.2.7}$$

para $i = 1, 2, \ldots, n-2$.

Isolando b_i na equação (5.2.6) podemos determiná-lo em termos dos valores conhecidos f_i, h_i e da incógnita c_i(o mesmo acontece com os coeficientes d_i, a partir da equação (5.2.5)),

$$b_i = \frac{f_{i+1} - f_i}{h_i} - \frac{h_i}{3}(2c_i + c_{i+1}), \tag{5.2.8}$$

para $i = 1, 2, \ldots, n-2$.

A substituição de b_i e b_{i-1} dados pela equação (5.2.8) na equação (5.2.7) com os índices deslocados de uma unidade, ou seja, $b_i = b_{i-1} + h_{i-1}(c_{i-1} + c_i)$, permite encontrar uma equação para os coeficientes c_i em termos dos valores conhecidos f_i e h_i:

$$h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} = 3\left(\frac{f_{i+1} - f_i}{h_i}\right) - 3\left(\frac{f_i - f_{i-1}}{h_{i-1}}\right), \tag{5.2.9}$$

para $i = 2, 3, \ldots, n-1$. A equação anterior define um sistema de equações lineares para as incógnitas c_i. Note que além dos coeficientes $c_1, c_2, \ldots, c_{n-1}$, o sistema envolve um coeficiente c_n que não está diretamente relacionado a algum dos $n-1$ polinômios s_i . Na realidade, c_n está relacionado às condições no extremo do intervalo de interpolação e sua determinação depende do tipo de spline que estamos construindo, se é um spline natural ou um spline completo (ou quase completo).

As $n-2$ equações (5.2.9) envolvem n variáveis (as incógnitas c_i), para que o sistema (tipicamente) tenha solução única devemos incluir as duas últimas equações que descrevem o comportamento do spline nos extremos do intervalo de interpolação. Vamos estudar inicialmente o caso do spline natural.

Spline natural

O spline natural deve satisfazer as condições $s''(x_1) = 0$ e $s''(x_n) = 0$, estas duas equações implicam respectivamente

$$c_1 = 0$$

e

$$2c_{n-1} + 6d_{n-1}h_{n-1} = 0. \tag{5.2.10}$$

A equação (5.2.10) implica em termos da equação para os coeficientes d_i (5.2.5) que $c_n = 0$.

Colecionando esses resultados temos então a seguinte situação: resolvendo o sistema de equações

$$\begin{cases} c_1 & = & 0 \\ h_{i-1}c_{i-1} + 2(h_{i-1}+h_i)c_i + h_i c_{i+1} & = & 3\left(\frac{f_{i+1}-f_i}{h_i}\right) - 3\left(\frac{f_i - f_{i-1}}{h_{i-1}}\right), \quad i = 2, \ldots, n-1 \\ c_n & = & 0 \end{cases}$$

encontramos o valor dos coeficientes c_i. A partir desses coeficientes determinamos o valor dos coeficientes b_i através das equações (5.2.8)

$$b_i = \frac{f_{i+1} - f_i}{h_i} - \frac{h_i}{3}(2c_i + c_{i+1}),$$

para $i = 1, 2, \ldots, n-1$; e o valor dos coeficientes d_i através da equações

$$d_i = \frac{c_{i+1} - c_i}{3h_i}, \tag{5.2.11}$$

para $i = 1, 2, \ldots, n-1$, obtida a partir de (5.2.5). Os coeficientes $a_i = f_i$ como já havíamos determinado anteriormente.

Spline completo ou quase completo

Nesse caso o spline deve satisfazer as condições $s'(x_1) = f'(x_1) \equiv f'_1$ e $s'(x_n) = f'(x_n) \equiv f'_n$.Para determinar o spline, f'_1 e f'_n devem ser valores conhecidos. As condições implicam respectivamente

$$b_1 = f'_1 \tag{5.2.12}$$

e

$$b_{n-1} + 2c_{n-1}h_{n-1} + 3d_{n-1}h_{n-1}^2 = f'_n. \tag{5.2.13}$$

Como os coeficientes b_i satisfazem a equação (5.2.8), a equação (5.2.12) implica

$$f_1' = b_1 = \frac{f_2 - f_1}{h_1} - \frac{h_1}{3}(2c_1 + c_2),$$

ou seja,

$$2h_1c_1 + h_1c_2 = 3\left(\frac{f_2 - f_1}{h_1}\right) - 3f_1'. \tag{5.2.14}$$

Da mesma forma, no caso da equação (5.2.13), as equações (5.2.8) e (5.2.11) implicam

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = -3\left(\frac{f_n - f_{n-1}}{h_{n-1}}\right) + 3f_n'. \tag{5.2.15}$$

Em resumo, devemos resolver o sistema formado pelas equações (5.2.9), (5.2.14) e (5.2.15)

$$\begin{cases} 2h_1c_1 + h_1c_2 & = & 3\left(\frac{f_2-f_1}{h_1}\right) - 3f_1' \\ h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_ic_{i+1} & = & 3\left(\frac{f_{i+1}-f_i}{h_i}\right) - 3\left(\frac{f_i-f_{i-1}}{h_{i-1}}\right), \quad i = 2, \ldots, n-1 \\ h_{n-1}c_{n-1} + 2h_{n-1}c_n & = & -3\left(\frac{f_n-f_{n-1}}{h_{n-1}}\right) + 3f_n' \end{cases}$$

e então determinar os coeficientes b_i e d_i através das equações (5.2.8) e (5.2.11):

$$b_i = \frac{f_{i+1} - f_i}{h_i} - \frac{h_i}{3}(2c_i + c_{i+1}),$$

$$d_i = \frac{c_{i+1} - c_i}{3h_i},$$

para $i = 1, 2, \ldots, n-1$. Naturalmente, os coeficientes $a_i = f_i$.

Exemplo 26: Vamos determinar a interpolação spline cúbica (natural e completo) para a função seno, no intervalo $[0, 2\pi]$ a partir de 5 pontos igualmente espaçados. Ou seja,

$$\begin{aligned}\{(x_i, f_i)\}_{i=1}^{5} &= \left\{\left(\frac{(i-1)}{2}\pi, \operatorname{sin}\left(\frac{(i-1)}{2}\pi\right)\right)\right\}_{i=1}^{5} \\ &= \left\{(0,0), \left(\frac{\pi}{2}, 1\right), (\pi, 0), \left(\frac{3\pi}{2}, -1\right), (2\pi, 0)\right\}\end{aligned}$$

Como o conjunto de pontos possui o mesmo espaçamento na coordenada x, então todos os h_i são iguais a $\dfrac{\pi}{2}$.

O spline cúbico é uma função $s(x)$ da forma

$$s(x) = \begin{cases} s_1(x), & 0 \leq x < \dfrac{\pi}{2} \\ s_2(x), & \dfrac{\pi}{2} \leq x < \pi \\ s_3(x), & \pi \leq x < \dfrac{3\pi}{2} \\ s_4(x), & \dfrac{3\pi}{2} \leq x \leq 2\pi \end{cases},$$

onde cada $s_i(x)$ é um polinômio cúbico

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

Quando se tratar de um spline natural então os coeficientes $c_i$são determinados através do sistema

$$\begin{cases} c_1 &= 0 \\ h_1c_1 + 2(h_1 + h_2)c_2 + h_2c_3 &= 3\left(\frac{f_3 - f_2}{h_2}\right) - 3\left(\frac{f_2 - f_1}{h_1}\right) \\ h_2c_2 + 2(h_2 + h_3)c_3 + h_3c_4 &= 3\left(\frac{f_4 - f_3}{h_3}\right) - 3\left(\frac{f_3 - f_2}{h_2}\right) \\ h_3c_3 + 2(h_3 + h_4)c_4 + h_4c_5 &= 3\left(\frac{f_5 - f_4}{h_4}\right) - 3\left(\frac{f_4 - f_3}{h_3}\right) \\ c_5 &= 0 \end{cases}. \tag{5.2.16}$$

Uma vez determinados os coeficientes c_i, os coeficientes b_i e d_i são funções deste:

$$b_i = \frac{f_{i+1} - f_i}{h_i} - \frac{h_i}{3}(2c_i + c_{i+1}),$$

$$d_i = \frac{c_{i+1} - c_i}{3h_i},$$

para $i = 1, 2, 3, 4$.

Neste exemplo $h_i = \dfrac{\pi}{2}$ para todo i. Substituindo os demais valores para f_i no sistema

(5.2.16) teremos

$$\begin{cases} c_1 & = & 0 \\ \frac{\pi}{2}c_1 + 2\pi c_2 + \frac{\pi}{2}c_3 & = & -\frac{12}{\pi} \\ \frac{\pi}{2}c_2 + 2\pi c_3 + \frac{\pi}{2}c_4 & = & 0 \\ \frac{\pi}{2}c_3 + 2\pi c_4 + \frac{\pi}{2}c_5 & = & \frac{12}{\pi} \\ c_5 & = & 0 \end{cases}$$

cuja solução é dada por

$$\begin{aligned} c_1 &= 0, \\ c_2 &= -\frac{6}{\pi^2}, \\ c_3 &= 0, \\ c_4 &= \frac{6}{\pi^2}, \\ c_5 &= 0. \end{aligned}$$

A partir dessa solução temos também que

$$\begin{aligned} b_1 &= \frac{3}{\pi}, \\ b_2 &= 0, \\ b_3 &= -\frac{3}{\pi}, \\ b_4 &= 0 \end{aligned}$$

e

$$\begin{aligned} d_1 &= -\frac{4}{\pi^3}, \\ d_2 &= \frac{4}{\pi^3}, \\ d_3 &= \frac{4}{\pi^3}, \\ d_4 &= -\frac{4}{\pi^3}. \end{aligned}$$

Como $a_i = f_i$ para $i = 1, 2, 3, 4$, o spline natural cúbico é dado então por

$$s(x) = \begin{cases} \frac{3}{\pi}x - \frac{4}{\pi^3}x^3, & 0 \le x < \frac{\pi}{2} \\ 1 - \frac{6}{\pi^2}\left(x - \frac{\pi}{2}\right)^2 + \frac{4}{\pi^3}\left(x - \frac{\pi}{2}\right)^3, & \frac{\pi}{2} \le x < \pi \\ -\frac{3}{\pi}(x - \pi) + \frac{4}{\pi^3}(x - \pi)^3, & \pi \le x < \frac{3\pi}{2} \\ -1 + \frac{6}{\pi^2}\left(x - \frac{3\pi}{2}\right)^2 - \frac{4}{\pi^3}\left(x - \frac{3\pi}{2}\right)^3, & \frac{3\pi}{2} \le x \le 2\pi \end{cases}.$$

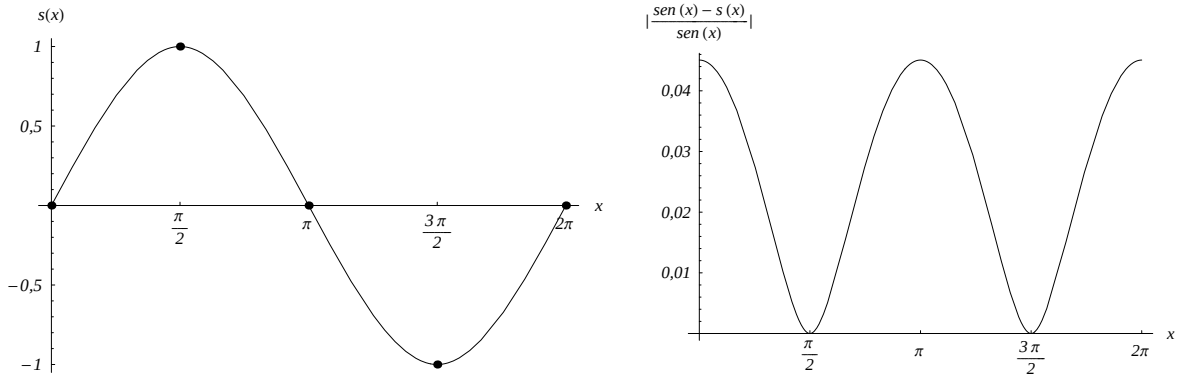


Figura 5.2.2: **a**) Interpolação spline cúbica (natural) da função seno em 5 pontos igualmente espaçados no intervalo $[0, 2\pi]$. **b**) Diferença entre a interpolação e a função seno. Note que o erro absoluto é menor ou igual a 0.05. O leitor mais atento observará que o gráfico para o erro relativo não se anula nos pontos pontos $x = 0$, $x = \pi$ e $x = 2\pi$. Acontece que nesses pontos o erro relativo não está definido pois o seno e sua diferença com relação a $s(x)$ se anulam. No entanto o limite existe e é indicado pelo gráfico.

O sistema para as constantes c_i no caso de um spline completo assume a forma

$$\begin{cases} 2h_1c_1 + h_1c_2 &= 3\left(\frac{f_2-f_1}{h_1}\right) - 3f_1' \\ h_1c_1 + 2(h_1+h_2)c_2 + h_2c_3 &= 3\left(\frac{f_3-f_2}{h_2}\right) - 3\left(\frac{f_2-f_1}{h_1}\right) \\ h_2c_2 + 2(h_2+h_3)c_3 + h_3c_4 &= 3\left(\frac{f_4-f_3}{h_3}\right) - 3\left(\frac{f_3-f_2}{h_2}\right) \\ h_3c_3 + 2(h_3+h_4)c_4 + h_4c_5 &= 3\left(\frac{f_5-f_4}{h_4}\right) - 3\left(\frac{f_4-f_3}{h_3}\right) \\ h_4c_4 + 2h_4c_5 &= -3\left(\frac{f_5-f_4}{h_4}\right) + 3f_5' \end{cases}.$$

Neste exemplo $h_i = \dfrac{\pi}{2}$ para todo i. Substituindo os demais valores para f_i no sistema acima teremos após algumas manipulações algébricas

$$\begin{cases} 2c_1 + c_2 &= \dfrac{12-6\pi}{\pi} \\ \dfrac{\pi}{2}c_1 + 2\pi c_2 + \dfrac{\pi}{2}c_3 &= -\dfrac{12}{\pi} \\ \dfrac{\pi}{2}c_2 + 2\pi c_3 + \dfrac{\pi}{2}c_4 &= 0 \\ \dfrac{\pi}{2}c_3 + 2\pi c_4 + \dfrac{\pi}{2}c_5 &= \dfrac{12}{\pi} \\ c_4 + 2c_5 &= -\left(\dfrac{12-6\pi}{\pi}\right) \end{cases}$$

cuja solução é dada por

$$\begin{aligned} c_1 &= -\frac{24}{7\pi^2}(-3+\pi), \\ c_2 &= \frac{6}{7\pi^2}(-10+\pi), \\ c_3 &= 0, \\ c_4 &= -\frac{6}{7\pi^2}(-10+\pi), \\ c_5 &= \frac{24}{7\pi^2}(-3+\pi). \end{aligned}$$

A partir dessa solução temos também que

$$\begin{aligned} b_1 &= 1, \\ b_2 &= -\frac{2}{7\pi}(-3+\pi), \\ b_3 &= \frac{-24+\pi}{7\pi}, \\ b_4 &= -\frac{2}{7\pi}(-3+\pi) \end{aligned}$$

e

$$\begin{aligned} d_1 &= \frac{4}{7\pi^3}(-22+5\pi), \\ d_2 &= -\frac{4}{7\pi^3}(-10+\pi), \\ d_3 &= \frac{4}{7\pi^3}(-10+\pi), \\ d_4 &= \frac{4}{7\pi^3}(-22+5\pi). \end{aligned}$$

Como $a_i = f_i$ para $i = 1, 2, 3, 4$, o spline cúbico completo é dado então por

$$s(x) = \begin{cases} s_1(x), & 0 \le x < \dfrac{\pi}{2} \\ s_2(x), & \dfrac{\pi}{2} \le x < \pi \\ s_3(x), & \pi \le x < \dfrac{3\pi}{2} \\ s_4(x), & \dfrac{3\pi}{2} \le x \le 2\pi \end{cases},$$

onde

$$\begin{aligned} s_1(x) &= x - \frac{24(-3+\pi)}{7\pi^2}x^2 + \frac{4(-22+5\pi)}{7\pi^3}x^3 \\ s_2(x) &= 1 - \frac{2(-3+\pi)}{7\pi}\left(x-\frac{\pi}{2}\right) + \frac{6(-10+\pi)}{7\pi^2}\left(x-\frac{\pi}{2}\right)^2 - \frac{4(-10+\pi)}{7\pi^3}\left(x-\frac{\pi}{2}\right)^3, \\ s_3(x) &= \frac{-24+\pi}{7\pi}(x-\pi) - \frac{4(-10+\pi)}{7\pi^3}(x-\pi)^3, \\ s_4(x) &= 1 - \frac{2(-3+\pi)}{7\pi}\left(x-\frac{3\pi}{2}\right) - \frac{6(-10+\pi)}{7\pi^2}\left(x-\frac{3\pi}{2}\right)^2 + \frac{4(-22+5\pi)}{7\pi^3}\left(x-\frac{3\pi}{2}\right)^3. \end{aligned}$$

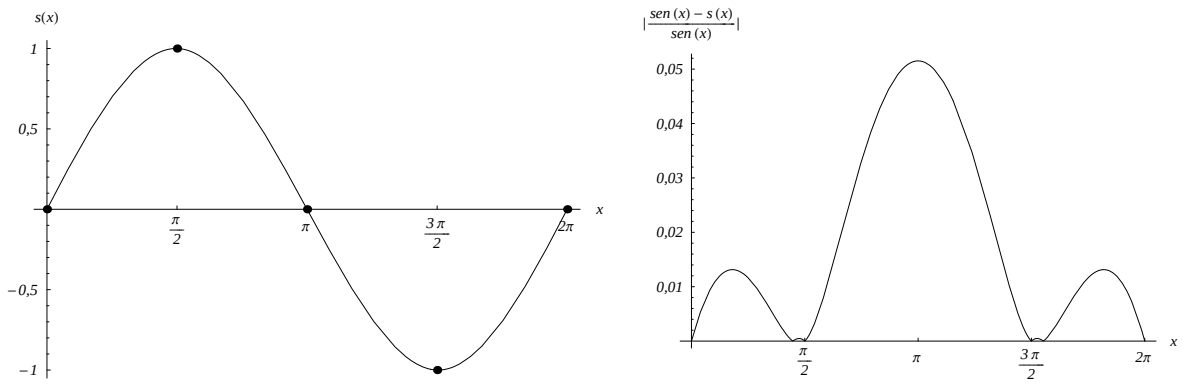


Figura 5.2.3: **a**) Interpolação por spline cúbico completo da função seno em 5 pontos igualmente espaçados no intervalo $[0, 2\pi]$. **b**) Diferença entre a interpolação e a função seno. Note que o erro absoluto é menor ou igual a 0,052.

Note que a exigência de que o spline possua a mesma derivada que a função seno nos pontos $x = 0$ e $x = 2\pi$ diminui o erro relativo na vizinhança desses pontos em quase $\frac{1}{4}$ do valor original enquanto que o erro relativo na metade do intervalo sofre um aumento muito discreto.

Exemplos

1) Construa o spline quase completo para o conjunto de pontos dado pela tabela abaixo e determine o intervalo no qual a curvatura, $\frac{|y''(x)|}{(1+y'(x)^2)^{3/2}}$, do spline assume o maior valor.

x	0	1	2	3	3,1	3,2	3,8	4,5
y	0	1,5	1,8	1,5	1,5	1	0,2	0,3

(Sugestão: trabalhe com os coeficientes do spline e a função `sp_val` e determine o intervalo a partir do gráfico para a curvatura).

```
// Coordenadas dos pontos de interpolação
xi=[0 1.0 2.0 3.0 3.1 3.2 3.8 4.5];
yi=[0 1.5 1.8 1.5 1.5 1.0 0.2 0.3];

// Derivadas numéricas nas extremidades
dy_e=(yi(2)-yi(1))/(xi(2)-xi(1)); // esquerda.
dy_d=(yi(8)-yi(7))/(xi(8)-xi(7)); // direita.

// Coeficientes dos polinômios no spline cúbico quase completo.
spl_y=spline3_interp([xi;yi],dy_e,dy_d);

// Gráfico do spline cúbico no intervalo [0,4.5]
// Vamos utilizar 300 pontos nos gráficos:
xvar=linspace(0,4.5,300);
y=sp_val(spl_y,xi,xvar); // valores de y(x) nos pontos xvar.
plot(xvar,y);
```

```
xgrid;

// Coeficientes dos polinômios no spline quadrático para y'
spl_dy=[spl_y(:,2) 2*spl_y(:,3) 3*spl_y(:,4) zeros(7,1)];
dy=sp_val(spl_dy,xi,xvar); // valores de y'(x) nos pontos xvar.

// Gráfico do spline quadrático para y'
scf(); // abre nova janela gráfica.
plot(xvar,dy);
xgrid;

// Coeficientes dos polinômios no spline linear para y''
spl_d2y=[2*spl_y(:,3) 6*spl_y(:,4) zeros(7,2)];
d2y=sp_val(spl_d2y,xi,xvar); // valores de y''(x) nos pontos xvar.

// Gráfico do spline linear para y''
scf(); // abre nova janela gráfica.
plot(xvar,d2y);
xgrid;

// Abre nova janela gráfica
scf();
// Gráfico da curvatura local, |y''|/((1+y'^2)^1.5)
plot(xvar,abs(d2y)./((1+dy^2)^1.5));
xgrid;
```

A partir do gráfico da curvatura, podemos verificar que a região de maior curvatura está no intervalo $(3; 3{,}1)$.

2) As matrizes `marchal` e `marcha2` contêm valores da relação entre velocidade e aceleração de um veículo para duas diferentes marchas em unidades SI. A velocidade ótima para a troca de marchas é dada pelo ponto no qual as curvas se cruzam. Neste caso, entre 22m/s e 26m/s. Construa um spline cúbico quase completo para as duas relações de marcha e determine uma aproximação para a velocidade ótima com quatro dígitos.

```
// relação vel. X acel. na 1ª marcha.
marchal=[5 10 14 20 22 26;6.5 7.4 7.9 7 6.6 5.1];
// Derivadas numéricas nas extremidades
dal_e=(7.4-6.5)/(10-5); // esquerda.
dal_d=(5.1-6.6)/(26-22); // direita.

// relação vel. X acel. na 2ª marcha.
marcha2=[5 10 14 18 22 26; 4.5 5.7 6.2 6.4 6.0 5.5];
// Derivadas numéricas nas extremidades
da2_e=(5.7-4.5)/(10-5); // esquerda.
da2_d=(5.5-6.0)/(26-22); // direita.

// Coeficientes dos polonômios nos splines cúbicos
spl_al=spline3_interp(marchal,dal_e,dal_d);
```

```
spl_a2=spline3_interp(marcha2,da2_e,da2_d);

// Gráfico dos splines cúbicos para marchal e marcha2
// Inicialmente desenharemos apenas os pontos de
// interpolação para marchal e marcha2
scf(); \\ abre nova janela gráfica.
plot(marchal(1,:),marchal(2,:),'.r');
plot(marcha2(1,:),marcha2(2,:),'.r');
// Vamos utilizar 300 pontos no intervalo [5,26] para
// desenhar as curvas dos splines.
v_var=linspace(5,26,300);
plot(v_var,sp_val(spl_a1,marcha(1,:),v_var),'-g');
plot(v_var,sp_val(spl_a2,marcha(2,:),v_var),'-g');
xgrid;

// As duas curvas se encontram em um ponto no intervalo
// (22,26). A velocidade de troca corresponde ao zero
// da função
function z=f(x)
z=spl_a1(5,1)-spl_a2(5,1)..
+(spl_a1(5,2)-spl_a2(5,2))*(x-22)..
+(spl_a1(5,3)-spl_a2(5,3))*(x-22)^2 ..
+(spl_a1(5,4)-spl_a2(5,4))*(x-22)^3
endfunction

v=zero_newraph(f,23.5,100);
```

A velocidade de troca ótima vale aproximadamente 24,56m/s.

5.3 Exercícios

1) (Aquecimento) Cheque, indiretamente, a exatidão das bibliotecas de funções de seu computador ou calculadora científica através da análise do comportamento das seguintes identidades nos valores de $x = i\frac{\pi}{20}$, para $i = 1, 2, \ldots, 9$.

1. $\text{sen}^2(x) + \cos^2(x) = 1$
2. $\text{sen}(2x) = 2\text{sen}(x)\cos(x)$
3. $\cos(x) = \text{sen}(x + \pi/2)$
4. $\exp(x)\exp(-x) = 1$
5. $\ln(e^x) = x$
6. $\sqrt{x}\sqrt{x} = x$

2) Utilize a seguinte tabela (com valores exatos até a precisão utilizada),

x	$\text{sen}(x)$	$\cos(x)$	$\cot(x)$
0,001	0,001000	1,000000	1000,0
0,002	0,002000	0,999998	499,999
0,003	0,003000	0,999996	333,332
0,004	0,004000	0,999992	249,999
0,005	0,00500	0,999988	199,998

para calcular $\cot(0{,}0015)$ com a maior precisão possível através de:

1. interpolação para $\cot(x)$.
2. interpolação de $\text{sen}(x)$ e $\cos(x)$.
3. estime o erro em 2). Dica: propagação de erros.
4. Explique a diferença entre os resultados em 1) e 2).

3) Compare os erros na aproximação das funções abaixo no intervalo $[0, 1]$ através de:
i) Expansão de Taylor em torno do ponto $x_0 = 0{,}5$
ii) Interpolação de Lagrange com pontos igualmente espaçados, com $x_1 = 0$ até $x_4 = 1$.
iii) Interpolação de Lagrange utilizando os pontos de Chebyshev.
Utilize sempre polinômios de 3º grau e compare os erros em $x = 0;\ 0{,}1;\ 0{,}2;\ \ldots;\ 1{,}0$.

1. $\text{sen}(2x)$
2. e^x
3. $\sqrt{x}$
4. $\dfrac{1}{1 + 25x^2}$

5. x^4

4) Seja a função $f(x) = (x - 0{,}2)(x - 0{,}3)e^{-(x+0{,}5)^2}$. Escolha um número suficiente de pontos distintos e construa um polinômio interpolante de terceiro grau no intervalo $[0{,}1; 0{,}4]$ para a função f. (Sugestão: utilize pontos igualmente espaçados.)

5) Encontre a interpolação *spline* cúbica (*spline* natural) para os dados abaixo

x	$f(x)$
-2	0
-1	1
0	2
1	1
2	0
3	1

6) Verifique se as seguintes funções são *splines*

1. $f(x) = \begin{cases} x & , \quad -1 \leq x < 0 \\ 2x & , \quad 0 \leq x < 1 \\ x+1 & , \quad 1 \leq x \leq 2 \end{cases}$

2. $f(x) = \begin{cases} x & , \quad -1 \leq x < 0 \\ 2x-1 & , \quad 0 \leq x < 1 \\ x+1 & , \quad 1 \leq x \leq 2 \end{cases}$

3. $f(x) = \begin{cases} 0 & , \quad -1 \leq x < 0 \\ x^2 & , \quad 0 \leq x < 1 \\ 2x-1 & , \quad 1 \leq x \leq 2 \end{cases}$

7) Determine os valores de a e b de forma que a seguinte função

$$f(x) = \begin{cases} x^3 + x & , \quad -1 \leq x < 0 \\ \\ ax^2 + bx & , \quad 0 \leq x \leq 1 \end{cases}$$

seja um *spline* cúbico.

8) Para quais valores de b, c e d a função $s : [-1, 1] \to \mathbb{R}$ é um *spline*? Quando será um *spline* natural? Justifique suas afirmações.

$$s(x) = \begin{cases} dx^3 - 9x^2 + bx & , \quad -1 \leq x < 0 \\ 3x^3 + cx^2 + x & , \quad 0 \leq x \leq 1 \end{cases}.$$

9) Considere os dados da tabela abaixo.

x_i	1.,0	1,1	1,3	1,5	1,9	2,3	2,4	2,7	2,9	3,0
$f(x_i)$	-0,9	-0,5	-0,1	0,3	0,4	0,5	1,0	1,3	1,5	1,9
$g(x_i)$	1,6	1,5	1,5	1,4	1,3	1,0	0,5	-0,1	-0,3	-0,5

Construa um polinômio interpolante para cada conjunto de dados e determine o valor de $x \in [1, 3]$ (arredondado para seis dígitos) no qual as curvas se encontram.

10) A relação entre ponto de congelamento e proporção (em peso) da mistura de água e glicerina é dada pela tabela abaixo

%glicerina (em peso)	10	20	30	40	50	60	70	80	90	100
temp. de congel. (ºC)	$-1{,}6$	$-4{,}8$	$-9{,}5$	$-15{,}5$	$-22{,}0$	$-33{,}6$	$-37{,}8$	$-19{,}2$	$-1{,}6$	17

Determine uma estimativa com três dígitos para a temperatura de congelamento de uma solução com 27% de glicerina (em peso) a partir da interpolação polinomial de quatro valores da tabela no intervalo de concentração de glicerina entre 10% e 40%.

11) Utilize os dados da tabela anterior e determine uma aproximação com três dígitos para o valor mínimo da temperatura de congelamento de uma mistura de água e glicerina. (Sugestão: construa o polinômio interpolante p e encontre a raiz real de p' no intervalo dado pela tabela).

12) A partir dos dados da tabela abaixo,

temperatura (ºC)	0	4	5	10	15
densidade (g/cm^3)	0,99984	0,99997	0,99996	0,99970	0,99910

construa dois polinômios interpolantes de grau 3, um adequado à temperatura de 3,5°C e outro à 12°C e os utilize para encontrar uma aproximação com cinco dígitos para o volume ocupado por uma determinada massa de água a 3,5°C se a 12°C a mesma massa ocupa um volume de 1L.

13) A expansão/contração linear de um objeto é obtida a partir da expressão

$$l(T) = l_0 + l_0 \int_{T_0}^{T} \alpha(\tau) d\tau,$$

onde $l(T)$ é o comprimento linear do objeto a uma temperatura T, l_0 é o comprimento desse mesmo objeto a uma temperatura T_0 e $\alpha(\tau)$ é o coeficiente de expansão térmica linear do objeto a uma temperatura τ. A tabela abaixo contém os valores para o coeficiente de expansão térmica linear de duas substâncias; um material cerâmico e o fluoreto de escândio (ScF_6):

T (K)	200	400	600	800	1000	1200	1400
$\alpha_{\text{cer}}(10^{-6}\text{K}^{-1})$	5,04	7,88	7,99	9,48	9,92	10,1	10,2
$\alpha_{\text{ScF}_6}(10^{-6}\text{K}^{-1})$	-9,69	-5,44	-2,82	-1,26	-0,238	0,578	1,87

Um determinado material, composto por uma fração r de material cerâmico e $(1-r)$ do fluoreto de escândio, possui a seguinte expressão para expansão linear em função da temperatura

$$l(T) = l_0 + l_0 \left(r \int_{T_0}^{T} \alpha_{\text{cer}}(\tau) d\tau + (1-r) \int_{T_0}^{T} \alpha_{\text{ScF}_6}(\tau) d\tau \right).$$

Determine o valor mínimo e máximo para a razão $l(T)/l_0$ quando $r = 0.3$, $T_0 = 350$ e T pertence ao intervalo $[200, 500]$. (Observação: utilize todos os dados no seus cálculos).

medskip

14) Os dados da tabela abaixo contém as posições de um corpo ao longo do tempo (em unidades SI).

t_i	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
x_i	0	4,97e-3	1,15e-2	1,59e-2	1,90e-2	2,24e-2	2,18e-2	1,48e-2	6,15e-3	1,45e-4	-7,46e-3
y_i	0	7,28e-5	4,52e-3	1,27e-2	2,25e-2	2,59e-2	2,30e-2	1,55e-2	2,15e-3	-1,40e-2	-2,74e-2

A partir das expressões dos splines cúbicos para cada coordenada, determine o maior valor (com três dígitos) que o módulo da velocidade

$$\sqrt{(x'(t))^2 + (y'(t))^2}$$

assume ao longo da trajetória.

15) A partir dos dados abaixo, construa um spline cúbico quase completo e determine uma aproximação com três dígitos para o valor t^* no qual o spline se anula.

t	0	0,1	0,18	0,27	0,36	0,44	0,53	0,62	0,71	0,8
$f(t)$	1	0,986	0,946	0,879	0,788	0,676	0,545	0,399	0,242	0,0785

16) Os dados da tabela abaixo contém as coordenadas de dois corpos ao longo do tempo.

t	0	0,1	0,18	0,27	0,36	0,44	0,53	0,62	0,71	0,8
x_1	0	0,174	0,342	0,5	0,643	0,766	0,866	0,94	0,985	1
y_1	1	0,98	0,921	0,824	0,695	0,537	0,358	0,165	-0,0349	-0,233
x_2	1,5	1,32	1,14	0,978	0,831	0,707	0,609	0,541	0,505	0,503
y_2	1	0,986	0,946	0,879	0,788	0,676	0,545	0,399	0,242	0,0785

Dado ainda que $\frac{dx_1}{dt}(0) = 2$, $\frac{dy_1}{dt}(0) = 0$, $\frac{dx_1}{dt}(0{,}8) = -0{,}0584$, $\frac{dy_1}{dt}(0{,}8) = -2{,}22$, $\frac{dx_2}{dt}(0) = -2{,}1$, $\frac{dy_2}{dt}(0) = 0$, $\frac{dx_2}{dt}(0{,}8) = 0{,}229$ e $\frac{dy_2}{dt}(0.8) = -1.9$; utilize a técnica de interpolação segmentada (splines cúbicos) para determinar uma aproximação com quatro dígitos para a menor distância entre os corpos ao longo das suas trajetórias.

6 Ajuste de mínimos quadrados

No capítulo anterior, foram desenvolvidas técnicas para fornecer uma expressão matemática para o comportamento de um conjunto de pontos (pares ordenados) $\{(x_i, f_i)\}_{i=1}^{n}$. A idéia consistia em prescrever uma função modelo dependente de uma quantidade de parâmetros e então determinar o valor dos mesmos exigindo que a função assuma os valores f_i em cada x_i. Essa exigência determina equações que os parâmetros devem satisfazer. Uma vez resolvidas as equações, a expressão matemática para o comportamento dos pontos estará definida.

A exigência de que a função modelo se iguale a f_i nos pontos x_i pode ser substituída por uma outra exigência menos estrita. Por exemplo, poderíamos estar interresados em limitar de alguma forma os erros cometidos na representação do comportamento dos pontos pelo modelo.

Seja então $\phi(x)$, uma função candidata a modelar o comportamento dos dados. Se relaxarmos a exigência de que ϕ é uma interpolação, em cada ponto x_i, a função pode não valer exatamente f_i. Nesse caso, haverá um resíduo $r_i = f_i - \phi(x_i)$. Agora, o objetivo é controlar o total desses resíduos. Para tanto, os parâmetros que definem ϕ devem ser tais que a totalidade desses resíduos seja a menor possível (mas sem que sejam todos iguais a zero necessariamente).

O coleção dos resíduos pode ser controlada a partir de uma séria de medidas. Por exemplo, $\sum_i |r_i|$, $\sum_i r_i^2$ ou ainda $\max_i |r_i|$. No entanto, se os erros que impedem uma fiel reprodução do comportamento dos pontos pela função ϕ possuirem a natureza de variáveis aleatórias descorrelacionadas, de média nula e mesma variância, é possível demonstrar[1] que a melhor escolha de parâmetros para ϕ é dada pelo mínimo valor que $\sum_i r_i^2$ assumir. Por essa razão, a tarefa de determinar a função ϕ que satisfaça esse critério é determinada ajuste de mínimos quadrados.

Assim, seja ϕ uma função de uma variável, definida a menos de um conjunto de m parâmetros $\{a_j\}_{j=1}^{m}$, $\phi\left(x; \{a_j\}_{j=1}^{m}\right)$. O ajuste de mínimos quadrados de ϕ ao conjunto de pontos $P = \{(x_i, f_i)\}_{i=1}^{n}$, consiste em determinar os valores $\left\{a_j^*\right\}_{j=1}^{m}$ que correspondam ao mínimo da função Q

$$Q_P(a_1, a_2, \ldots, a_m) := \sum_{i=1}^{n} \left(f_i - \phi(x_i; a_1, a_2, \ldots, a_m)\right)^2 .$$

Se a dependência de ϕ nos parâmetros a_j não for linear, a tarefa de determinar o mínimo de Q_P pode ser consideravelmente complexa. Nos mínimos locais, as derivadas parciais de Q_P se anulam, portanto, o ponto de partida consiste em montar as equações a partir dessa exigência. No entanto, os mínimos locais são um dos tipos de ponto crítico associados à solução dessas equações. Os outros são os pontos de máximo e os pontos de sela. Assim, além de resolver essas equações é necessário conhecer o comportamento da matriz hessiana (nesse caso, é necessário conhecer seus autovalores) no ponto crítico para então confirmar se esse ponto crítico se trata de um mínimo

[1] Esse resultado é um teorema em estatística, conhecido como teorema de Gauss-Markov.

local. A tarefa seguinte é escolher o menor dos mínimos. Se for levado em consideração que podem existir infinitas soluções, o quadro já não é tão animador.

Felizmente, de maneira análoga à que ocorre nos problemas de interpolação, se a dependência de ϕ nos parâmetros for linear, essas dificuldades são prontamente eliminadas. Esse é o caso do ajuste de mínimos quadrados linear.

6.1 Ajuste de mínimos quadrados linear

Se a função que se deseja ajustar for uma combinação linear de m funções conhecidas e linearmente independentes, a tarefa de se determinar um conjunto de parâmetros que minimize a soma do quadrado dos resíduos consiste em resolver um sistema de equações lineares.

Seja portanto, a combinação linear formada por um conjunto linearmente independente formado por m funções conhecidas $\{\varphi_j(x)\}_{j=1}^{m}$:

$$\phi(x) = \sum_{j=1}^{m} a_j \varphi_j(x).$$

Uma vez fixados os dados e o conjunto de funções, a soma quadrática dos resíduos[2] depende apenas dos coeficientes da combinação linear,

$$Q\left(a_1, a_2, \ldots, a_m\right) := \sum_{i=1}^{n} \left(f_i - \phi(x_i)\right)^2 = \sum_{i=1}^{n} \left(f_i - \sum_{j=1}^{m} a_j \varphi_j(x_i) \right)^2 . \tag{6.1.1}$$

Nesse caso, podemos notar que em termos dos coeficientes a_j, Q representa um parabolóide imerso no $\mathbb{R}^{m+1}$. Basta agora verificar a natureza do ponto crítico dessa superfície. Seria desejável que o ponto crítico correspondesse ao mínimo de Q, mas antes disso é necessário excluir a possibilidade de que ele seja um ponto de sela.

O ponto crítico corresponde à solução do sistema de equações

$$\begin{cases} \dfrac{\partial Q}{\partial a_1} &= 0 \\ \dfrac{\partial Q}{\partial a_2} &= 0 \\ & \vdots \\ \dfrac{\partial Q}{\partial a_m} &= 0 \end{cases}$$

[2] Daqui em diante, deixará de ser utilizado o subescrito P na definição da função Q.

A forma da derivada parcial de Q com respeito ao coeficiente a_k corresponde a

$$\begin{aligned}
\frac{\partial Q}{\partial a_k} &= \frac{\partial}{\partial a_k}\left(\sum_{i=1}^{n}\left(f_i - \sum_{j=1}^{m} a_j\varphi_j(x_i)\right)^2\right) \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial a_k}\left(\left(f_i - \sum_{j=1}^{m} a_j\varphi_j(x_i)\right)^2\right) \\
&= \sum_{i=1}^{n} 2\left(f_i - \sum_{j=1}^{m} a_j\varphi_j(x_i)\right)\frac{\partial}{\partial a_k}\left(f_i - \sum_{j=1}^{m} a_j\varphi_j(x_i)\right) \\
&= -2\sum_{i=1}^{n}\left(f_i - \sum_{j=1}^{m} a_j\varphi_j(x_i)\right)\varphi_k(x_i) \\
&= -2\left(\sum_{i=1}^{n} f_i\varphi_k(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} a_j\varphi_j(x_i)\varphi_k(x_i)\right), \qquad (6.1.2)
\end{aligned}$$

onde na passagem da primeira para a segunda linha, foi utilizado o fato de que a derivação é uma operação linear (ou seja, a derivada da soma de várias funções é igual a soma de suas derivadas). Assim a k-ésima equação do sistema pode ser escrita como

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_j\varphi_j(x_i)\varphi_k(x_i) = \sum_{i=1}^{n} f_i\varphi_k(x_i). \qquad (6.1.3)$$

Nesse ponto já podemos identificar claramente que o sistema de equações é linear nos coeficientes a_j. No entanto é útil exprimir o sistema de equações na forma matricial. Antes de fazê-lo, vale relembrar

Observação 6.1.1 (Multiplicação matricial). *Sejam duas matrizes $A_{p\times q}$ e $B_{q\times r}$ de componentes $a_{i,j} = (A)_{i,j}$ e $b_{i,j} = (B)_{i,j}$. Então, o elemento de índices i, j do produto $A \times B$ é dado pelo somatório (note que a ordem é importante)*

$$(A \times B)_{i,j} = \sum_{k=1}^{q} a_{i,k}b_{k,j},$$

para $i = 1, 2, \ldots, p$ e $j = 1, 2, \ldots, r$.

A partir dessa observação, é possível verificar que o lado esquerdo de (6.1.3) corresponde a dois produtos matriciais e o lado direito a um produto matricial. Sejam as matrizes Φ, $\mathbf{a}$ e $\mathbf{f}$, cujas componentes são dadas por

$$(\Phi)_{i,j} := \varphi_j(x_i), \qquad (\mathbf{a})_{j,1} = a_j \qquad \text{e} \qquad (\mathbf{f})_{i,1} = f_i.$$

O lado direito de (6.1.3) pode ser reescrito como

$$\begin{aligned}\sum_{i=1}^{n} f_i \varphi_k(x_i) &= \\ &= \sum_{i=1}^{n} (\mathbf{f})_{i,1} (\Phi)_{i,k} \\ &= \sum_{i=1}^{n} (\mathbf{f})_{i,1} \left(\Phi^T\right)_{k,i} \\ &= \left(\Phi^T \mathbf{f}\right)_{k,1},\end{aligned}$$

seguindo o mesmo desenvolvimento, o lado esquerdo pode ser reescrito como

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} a_j \varphi_j(x_i)\varphi_k(x_i) &= \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} (\mathbf{a})_{j,1} (\Phi)_{i,j} (\Phi)_{i,k} \\ &= \sum_{i=1}^{n} (\Phi\mathbf{a})_{i,1} (\Phi)_{i,k} \\ &= \left(\Phi^T \Phi \mathbf{a}\right)_{k,1}.\end{aligned}$$

Ou seja, a equação (6.1.3) corresponde à k-ésima linha da equação matricial

$$\Phi^T \Phi \mathbf{a} = \Phi^T \mathbf{f}, \tag{6.1.4}$$

onde

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_m(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_m(x_n) \end{pmatrix}_{n\times m},$$

$\mathbf{a} = (a_1\, a_2\, \ldots\, a_m)^T$ e $\mathbf{f} = (f_1\, f_2\, \ldots\, f_n)^T$.

Agora falta apenas garantir que o ponto crítico, solução do sistema (6.1.4), corresponde a um mínimo de Q. Para tanto, será necessário obter as componentes da matriz hessiana de Q. Tomando a derivada parcial de (6.1.2) com relação ao coeficiente a_l

$$\frac{\partial^2 Q}{\partial a_l \partial a_k} = 2\sum_{i=1}^{n} \varphi_l(x_i)\varphi_k(x_i) = 2\sum_{i=1}^{n} (\Phi)_{i,l} (\Phi)_{i,k} = 2\sum_{i=1}^{n} \left(\Phi^T\right)_{l,i} (\Phi)_{i,k} = 2\left(\Phi^T\Phi\right)_{l,k}.$$

Isto permite concluir que a matriz hessiana é igual $2\Phi^T\Phi$. Neste caso, por ser igual ao produto da transposta de uma matriz por ela mesma, $2\Phi^T\Phi$ é uma matriz positiva definida. Isto equivale a dizer que os autovalores da matriz hessiana de Q são reais positivos. Consequentemente, o ponto

crítico é um ponto de mínimo.

As equações dadas pelo sistema (6.1.4) são denominadas *equações normais*. Essa nomenclatura se deve ao fato do sistema admitir a forma

$$\Phi^T(\Phi\mathbf{a}-\mathbf{f})=\mathbf{0}_{m,1}.$$

O vetor entre parênteses, $\Phi\mathbf{a}-\mathbf{f}$ é o vetor cujas componentes são dadas pelos resíduos da aproximação e, segundo a equação anterior esse vetor é normal (ortogonal) aos vetores formados pelos elementos das linhas da matriz Φ^T que são da forma $(\varphi_j(x_1)\ \varphi_j(x_2)\ \ldots\ \varphi_j(x_n))$, para $j=1,2,\ldots,m$.

Exemplo 27: Seja o conjunto de pontos $\{(x_i,f_i)\}_{i=1}^{5}$:

$$\{(-2,0),(-1,1),(0,2),(1,1),(2,0)\}$$

vamos determinar o ajuste de mínimos quadrados para a seguinte combinação linear, $\varphi(x)=a_1+a_2e^x+a_3e^{-x}$, onde $\varphi_1(x):=1$, $\varphi_2(x):=e^x$ e $\varphi_3(x)=e^{-x}$.

Os coeficientes do ajuste são solução do seguinte sistema na representação matricial

$$\Phi^T\Phi\mathbf{a}=\Phi^T\mathbf{f},$$

onde a matriz Φ é dada por

$$\Phi=\begin{pmatrix}\varphi_1(x_1)&\varphi_2(x_1)&\varphi_3(x_1)\\ \varphi_1(x_2)&\varphi_2(x_2)&\varphi_3(x_2)\\ \varphi_1(x_3)&\varphi_2(x_3)&\varphi_3(x_3)\\ \varphi_1(x_4)&\varphi_2(x_4)&\varphi_3(x_4)\\ \varphi_1(x_5)&\varphi_2(x_5)&\varphi_3(x_5)\end{pmatrix}=\begin{pmatrix}1&e^{-2}&e^2\\ 1&e^{-1}&e\\ 1&1&1\\ 1&e&e^{-1}\\ 1&e^2&e^{-2}\end{pmatrix},$$

portanto

$$\Phi^T\Phi=\begin{pmatrix}5{,}00000&11{,}6106&11{,}6106\\ 11{,}6106&63{,}1409&5{,}00000\\ 11{,}6106&5{,}0000&63{,}1409\end{pmatrix}.$$

O vetor de constantes $\Phi^T\mathbf{f}=\begin{pmatrix}4{,}00000\\ 5{,}08616\\ 5{,}08616\end{pmatrix}$ e o vetor de incógnitas $\mathbf{a}=\begin{pmatrix}a_1\\ a_2\\ a_3\end{pmatrix}$ compõe o sistema que possui matriz completa

$$\begin{pmatrix}5{,}00000&11{,}6106&11{,}6106&4{,}00000\\ 11{,}6106&63{,}1409&5{,}00000&5{,}08616\\ 11{,}6106&5{,}00000&63{,}1409&5{,}08616\end{pmatrix}$$

e solução

$$\mathbf{a} = \begin{pmatrix} 2{,}17256 \\ -0{,}295542 \\ -0{,}295542 \end{pmatrix}.$$

Assim, o ajuste toma a forma $\varphi(x) = 2{,}17256 - 0{,}295542\,(e^x + e^{-x})$.

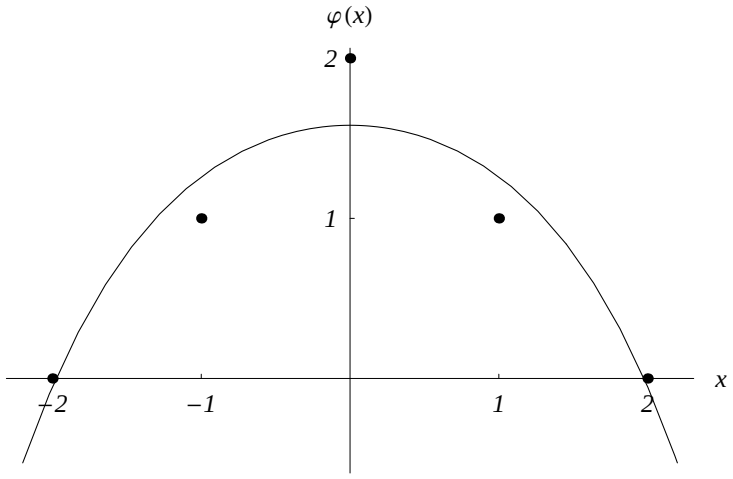


Figura 6.1.1: Ajuste de mínimos quadrados – combinação linear

Exemplo 28: (**Resolvendo um ajuste no Scilab**)

```
// Criação de um conjunto de dados para ajuste
// de mínimos quadrados. Esta primeira parte
// possui a função de criar os dados para o ajuste.

// inicialização do gerador de variáveis aleatórias.
rand('seed',1);

// função modelo
function y=f(x)
y=1.34+3.2*sin(%pi*x)+1.5*x^2;
endfunction

// Os dados deve estar na forma matriz coluna.
xp=linspace(0,5)';
// Adição de um termo aleatório entre -1.5 e 1.5
// ao valor da função modelo.
yp=f(xp)+1.5*(2*rand(xp)-1);

// Gráfico dos pontos para o ajuste.
plot(xp,yp,'.r');
xgrid
// Final do código que cria o conjunto de dados.
////////////////////////////////////////////////////////////

// Ajuste de mínimos quadrados.
```

```
// Desejamos determinar a combinação linear das funções
// 1, sin(%pi*x) e x^2 que corresponde ao ajuste pelos
// dados (xp,yp). Ou seja, phi1=1, phi2=sin(%pi*x) e
// phi3=x^2.

// Criação da matriz Phi
Phi=[ones(xp) sin(%pi*xp) xp^2];
// Solução do sistema
a=(Phi'*Phi)\(Phi'*yp);

// Gráfico do ajuste que acabamos de determinar.
// Vamos trabalhar com 300 pontos entre 0 e 5
x_var=linspace(0,5,300);
// Definição da função ajustada
// Note que a variável ``a'' contém os coeficientes
// do ajuste.
function z=f_aj(x)
z=a(1)+a(2)*sin(%pi*x)+a(3)*x^2;
endfunction

plot(x_var,f_aj(x_var));
```

Ajuste de polinômios

Um caso particular de importância prática é o ajuste de mínimos quadrados de um polinômio de grau m. Como qualquer polinômio pode ser escrito como uma combinação linear de monômios linearmente independentes, a estrutura matricial que já estudamos é mantida.

Seja então $p(x)$ um polinômio

$$p(x) = a_0 + a_1 x + \ldots + a_m x^m$$

e o conjunto de dados $\{(x_i, f_i)\}_{i=1}^{n}$. Os coeficientes do ajuste de mínimos quadrados desse polinômio pelos dados são solução do sistema de equações lineares

$$X^T X \mathbf{a} = X^T \mathbf{f}, \tag{6.1.5}$$

onde

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{pmatrix}, \tag{6.1.6}$$

$\mathbf{a} = (a_0\, a_1\, \ldots\, a_m)^T$ e $\mathbf{f} = (f_0\, f_1\, \ldots\, f_n)^T$.

Exemplo 29: Seja o conjunto de pontos $\{(x_i, f_i)\}_{i=1}^{5}$:

$$\{(-2,0), (-1,1), (0,2), (1,1), (2,0)\}.$$

Vamos determinar o ajuste de mínimos quadrados para um polinômio de segundo grau $p(x) = a_0 + a_1 x + a_2 x^2$ a esses dados.

Os coeficientes do polinômio são a solução do seguinte sistema na representação matricial

$$X^T X \mathbf{a} = X^T \mathbf{f},$$

onde a matriz X é dada por

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix},$$

portanto

$$X^T X = \begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix}.$$

O vetor de constantes $\mathbf{f} = (f_1\, f_2\, f_3\, f_4\, f_5)^T$ e o vetor de incógnitas $\mathbf{a} = (a_0\, a_1\, a_2)^T$ compõe o sistema que possui matriz completa

$$\begin{pmatrix} 5 & 0 & 10 & 4 \\ 0 & 10 & 0 & 0 \\ 10 & 0 & 34 & 2 \end{pmatrix}$$

e solução

$$\mathbf{a} = \begin{pmatrix} 58/35 \\ 0 \\ -3/7 \end{pmatrix}.$$

Assim, o polinômio que ajusta os dados é $p(x) = \dfrac{58}{35} - \dfrac{3}{7}x^2$.

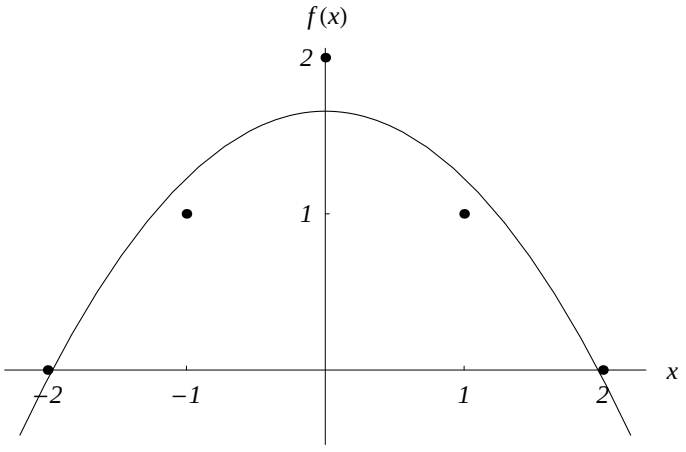


Figura 6.1.2: Ajuste de mínimos quadrados – ajuste polinomial

Problemas de condicionamento no método de ajuste de funções pelo método dos mínimos quadrados

De modo geral, ao aplicarmos o método de ajuste de mínimos quadrados com um polinômio de grau elevado (a partir de grau 7 nos sistemas que trabalham com aritmética de ponto flutuante de 64bits), a tarefa de resolver o sistema de equações normais (6.1.5) é muito dificultada por erros de arredondamento. A dificuldade está relacionada ao fato de que as matrizes da forma $X^T X$ presentes no sistema (6.1.5) são mal condicionadas. Essa propriedade independe dos valores f_i no conjunto de n pontos $\{(x_i, f_i)\}_{i=1}^n$ utilizados no ajuste. Note que a matriz $X^T X$ depende apenas dos valores x_i e do grau m do polinômio ajustado.

Como exemplo, no caso em que os n valores x_i são igualmente espaçados entre 0 e 1 é possível demonstrar[3] que $X^T X$ é aproximadamente igual a matriz $n\mathfrak{H}$, onde $\mathfrak{H}$ é a matriz de Hilbert[4] de ordem $m+1$, uma matriz mal condicionada.

Uma maneira de contornar as dificuldades introduzidas pelo condicionamento da matriz $X^T X$ e, ainda assim, realizar o ajuste de mínimos quadrados para um polinômio de ordem grande, consiste em utilizar um conjunto de polinômios $\varphi_i(x)$ construído de maneira que a matriz $\Phi^T \Phi$ presente no sistema de equações normais (6.1.4) não possua problemas de condicionamento. Como veremos adiante, esse objetivo é alcançado se o conjunto de funções $\{\varphi_i(x)\}_i$ for um conjunto de *funções ortogonais*.

Ajuste de funções ortogonais

Antes de mais nada, será necessário definir o conceito de produto interno para funções definidas em um conjunto discreto de pontos.

Definição 6.1.2 (produto interno discreto). *Seja o conjunto finito de pontos $X = \{x_i\}_{i=1}^n$ e duas funções f e g definidas sobre X. O produto interno discreto entre f e g, simbolizado pela expressão $(f,g)_X$ é definido como*

$$(f,g)_X = \sum_{i=1}^{n} f(x_i) g(x_i)$$

Definição 6.1.3 (funções ortogonais). *Dadas duas funções f e g, definidas em conjunto discreto finito X, dizemos que as mesmas são ortogonais se*

$$(f,g)_X = 0.$$

Em particular, um conjunto de funções $\{\varphi_i(x)\}_{i=1}^m$ definidas nos pontos do conjunto X é um *sistema ortogonal* se e somente se, para quaisquer $1 \leq i, j \leq m$,

$$\begin{aligned} (\varphi_i, \varphi_j)_X &= 0, \quad \text{se } i \neq j, \\ (\varphi_i, \varphi_j)_X &\neq 0, \quad \text{se } i = j. \end{aligned}$$

[3] Veja a demonstração na referência:
Yakowitz, S.; Szidarovszky, F. *An Introduction to Numerical Computation*. Macmillan Pub. Company. (1986).

[4] A matriz de Hilbert $\mathfrak{H}$ possui coeficientes $(\mathfrak{H})_{i,j} = \dfrac{1}{i+j-1}$. O condicionamento dessa matriz cresce exponencialmente com a ordem.

Se o ajuste de mínimos quadrados a partir dados $\{(x_i, f_i)\}_{i=1}^n$ é realizado para uma combinação linear

$$\sum_{i=1}^{m} a_i \varphi_i(x), \tag{6.1.7}$$

onde as funções $\varphi_i(x)$ são elementos de um sistema ortogonal, então a matriz $\Phi^T\Phi$ é uma matriz diagonal. Isto simplifica a tarefa de resolver numericamente o sistema (6.1.4).

Seja s_{ij} um coeficiente da matriz $S = \Phi^T\Phi$ onde $\{\varphi_i(x)\}_{i=1}^m$ é um sistema ortogonal. A partir da definição da matriz Φ, temos que

$$\begin{aligned} s_{ij} &= \sum_{k=1}^{n} \varphi_i(x_k)\varphi_j(x_k) \\ &= (\varphi_i, \varphi_j)_X, \end{aligned}$$

como as funções φ_i fazem parte de um sistema ortogonal, então

$$s_{ij} = \begin{cases} 0 & , \text{se } i \neq j \\ (\varphi_i, \varphi_i)_X & , \text{se } i = j \end{cases},$$

ou seja, $S = \Phi^T\Phi$ é uma matriz diagonal. Nesse caso, a solução do sistema de equações normais pode ser obtida a um baixo custo computacional e com erros de arredondamento controláveis através da inversão de S:

$$\mathbf{a} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{f}.$$

como

$$\left(\Phi^T\Phi\right)^{-1} = \begin{pmatrix} \frac{1}{(\varphi_1,\varphi_1)_X} & 0 & \cdots & 0 \\ 0 & \frac{1}{(\varphi_2,\varphi_{2,})_X} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{(\varphi_m,\varphi_m)_X} \end{pmatrix},$$

temos então que os coeficientes do ajuste[5] (6.1.7) são dados por

$$a_j = \frac{(f, \varphi_j)_X}{(\varphi_j, \varphi_j)_X}.$$

Portanto, já que x_i e f_i (ou $f(x_i)$) são dados de entrada para o ajuste, a determinação dos coeficientes a_i depende apenas da tarefa de encontrar um conjunto de funções φ_i que seja um sistema ortogonal. Ou seja, dado um conjunto linearmente independente de funções, basta realizar o processo de ortonormalização de Gram-Schmidt para construir a base de funções ortonormais.

Se φ_i forem um polinômios de grau i, então um sistema de bases ortogonais é construído[6] a partir de relações de recorrência . De acordo com essa construção, os polinômios φ_i devem

[5] Lembre que os dados para ajuste são $\{(x_i, f_i)\}_{i=1}^n$, onde o conjunto $X = \{x_i\}_{i=1}^n$ é utilizado para definir o produto interno $(\cdot,\cdot)_X$. Como temos estudado até aqui, $f_i \equiv f(x_i)$ se f for conhecida, caso contrário f_i é um dado de entrada no problema de ajuste.

[6] Veja os detalhes em:
Ralston, A.; Rabinowitz, P. *A First Course in Numerical Analysis*. McGraw-Hill.(1978).

satisfazer:

$$\varphi_i(x) = (x - b_i)\varphi_{i-1}(x) - c_i\varphi_{i-2}(x),$$

onde

$$\begin{aligned} b_i &= \frac{(x\varphi_{i-1}(x), \varphi_{i-1}(x))_X}{(\varphi_{i-1}(x), \varphi_{i-1}(x))_X}, \quad i = 1, 2, \ldots \\ c_i &= \frac{(x\varphi_{i-1}(x), \varphi_{i-2}(x))_X}{(\varphi_{i-2}(x), \varphi_{i-2}(x))_X}, \quad i = 2, 3, \ldots \end{aligned}$$

e $c_1 = 0$. Assim, construímos recursivamente os polinômios a partir da escolha $\varphi_0(x) \equiv 1$ e $\varphi_{-1}(x) \equiv 0$.

6.2 Ajustes linearizados

Existem situações nas quais o comportamento dos dados que se deseja modelar não pode ser convenientemente descrito a partir de uma função modelo na forma de uma combinação linear mas sim por uma função como por exemplo, a exponencial, a lei de potência, a função gaussiana ou por produtos entre quaisquer uma dessas funções. Os parâmetros de ajuste que minimizam a soma do quadrado dos resíduos nessas funções fazem parte da solução de sistemas de equações não lineares. Ainda assim, funções modelo com essa característica possuem grande importância em problemas aplicados e naturalmente seria desejável que houvesse ao menos uma forma aproximada de ajuste que pudesse ser obtida sem a necessidade de resolver sistemas não lineares. No caso dos exemplos supracitados, existe uma transformação que os leva em uma combinação linear, a transformação pelo logaritmo.

Sejam as funções

$$\begin{aligned} \phi_{\text{exp}}(x) &= a_1 e^{a_2 x}, \\ \phi_{\text{pot}}(x) &= a_1 x^{a_2}, \\ \phi_{\text{pot.exp}}(x) &= a_1 x^{a_2} e^{a_3 x}, \\ \phi_{\text{g}}(x) &= a_1 e^{a_3 (x - a_2)^2}. \end{aligned}$$

Ao aplicar o logaritmo nos dois lados das expressões, o resultado à direita são combinações lineares de funções na variável "x". Para diferenciá-las das funções originais, adicionamos o sinal

tipográfico "~",

$$\begin{aligned}
\tilde{\phi}_{\text{exp}}(x) &= \tilde{a}_1 + \tilde{a}_2 x. && \text{Onde } \tilde{a_1} = \ln a_1 \text{ e } \tilde{a}_2 = a_2.\\
\tilde{\phi}_{\text{exp}}(x) &= \tilde{a}_1 + \tilde{a}_2 \ln x. && \text{Onde } \tilde{a_1} = \ln a_1 \text{ e } \tilde{a}_2 = a_2.\\
\tilde{\phi}_{\text{pot.exp}}(x) &= \tilde{a}_1 + \tilde{a}_2 \ln x + \tilde{a}_3 x. && \text{Onde } \tilde{a_1} = \ln a_1,\ \tilde{a}_2 = a_2 \text{ e } \tilde{a}_3 = a_3.\\
\tilde{\phi}_{\text{g}}(x) &= \tilde{a}_1 + \tilde{a}_2 x + \tilde{a}_3 x^2. && \text{Onde } \tilde{a_1} = a_3 a_2^2 + \ln a_1,\ \tilde{a}_2 = 2a_2 a_3 \text{ e } \tilde{a}_3 = a_3.
\end{aligned}$$

Se desejamos ajustar uma combinação linear $\tilde{\phi} := \ln \phi$ ao conjunto de dados $\{(x_i, f_i)\}_{i=1}^n$ cujo comportamento é próximo ao da função ϕ, então, a solução é obtida como um problema de ajuste linear da função $\tilde{\phi}$ ao conjunto de dados$\{(x_i, \ln f_i)\}_{i=1}^n$. Assim, o problema original é substituído por um outro mais elementar, o problema de determinar o ajuste de mínimos quadrados de uma combinação linear de funções ao conjunto de dados $\{(x_i, \ln f_i)\}_{i=1}^n$.

Observação 6.2.1. *Os coeficientes ajustados a partir do problema linearizado não são iguais aos obtidos através do ajuste de mínimos quadrados não linear. Porém, se o comportamento dos dados for próximo ao da função que desejamos ajustar, os coeficientes obtidos a partir da linearização são próximos dos coeficientes para o ajuste não linear.*

Exemplo 30: Vamos realizar o ajuste dos pontos $\{(0{,}1; 0{,}01), (0{,}2; 0{,}063), (0{,}5; 0{,}59), (0{,}7; 1{,}5), (1{,}0; 3{,}6)\}$ à função $\theta(x) = a_1 x^{a_2}$. Vamos tomar o logaritmo das duas coordenadas dos pontos, o resultado é o novo conjunto de pontos $\{(x_i, \ln \theta_i)\}_{i=1}^n = \{(x_i, f_i)\}_{i=1}^n$: $\{(0{,}1; -4{,}60517), (0{,}2; -2{,}76462), (0{,}5; -0{,}527633), (0{,}7; 0{,}405465), (1{,}0; 1{,}28093)\}$.

Como

$$\begin{aligned}
\ln \theta &= \ln a_1 + a_2 x\\
&= \tilde{a}_1 + a_2 x,
\end{aligned}$$

ou seja, $\varphi_1(x) = 1$ e $\varphi_2(x) = \ln x$. Devemos resolver o sistema

$$\Phi^T \Phi \mathbf{a} = \Phi^T \mathbf{f},$$

onde, de acordo com a notação de produto interno,

$$\Phi^T \Phi = \begin{pmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^5 1 & \sum_{i=1}^5 \ln x_i \\ \sum_{i=1}^5 \ln x_i & \sum_{i=1}^5 (\ln x_i)^2 \end{pmatrix},$$

$$\Phi^T \mathbf{f} = \begin{pmatrix} (\mathbf{f}, \varphi_1) \\ (\mathbf{f}, \varphi_2) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^5 f_i \\ \sum_{i=1}^5 f_i \ln x_i \end{pmatrix}$$

e

$$\mathbf{a} = \begin{pmatrix} \tilde{a}_1 \\ a_2 \end{pmatrix}.$$

A solução é dada por

$$\tilde{a_1} = 1{,}28619 \qquad \text{e} \qquad a_2 = 2{,}54784.$$

Como $a_1 = e^{\tilde{a_1}}$, obtemos o ajuste

$$\theta(x) = 3{,}61897\, x^{2{,}54784}.$$

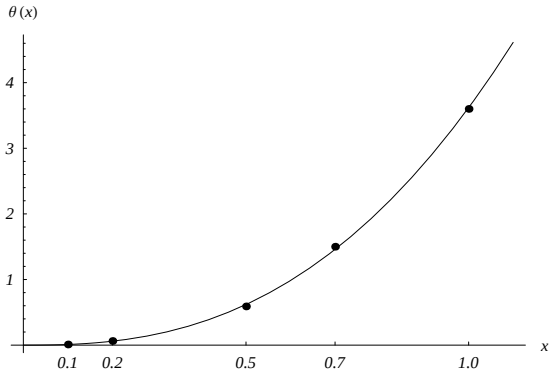


Figura 6.2.1: Ajuste de mínimos quadrados – ajuste não linear

6.3 Exercícios

1) Os dados da tabela abaixo contêm as magnitudes em escala Richter e o número de ocorrências de terremotos de magnitude maior ou igual a m em uma dada região e em um dado intervalo de tempo. Realize um ajuste de mínimos quadrados para determinar os coeficientes (com três dígitos) a e b da lei de Gutenberg-Richter,

$$N(m) = 10^{a-bm}$$

para a região onde os dados foram colhidos. (Sugestão: lembre que $\log_{10} N(m) = a - b\,m$ é uma combinação linear das funções 1 e $-m$).

m	4,0	4,2	4,4	4,6	4,8	5.0	5,2	5,4	5,6	5,8	6
$N(m)$	375	219	166	106	69	39	23	18	11	7	5

2) Os dados da tabela abaixo

t(h)	1	1,25	1,5	1,75	2	2,25	2,5	2,75	3
c(ng/ml)	6,28	4,74	3,29	2,57	1,88	1,56	0,889	0,525	0,375e

contém a concentração plasmática (em uma unidade padrão) de uma determinada substância no organismo de um ser vivo horas após sua ingestão. Utilize esses dados para ajustar a função $c(t) = a_1 e^{a_2 t}$ e determinar uma aproximação com quatro dígitos para o tempo no qual a concentração corresponda a 25% da concentração inicial.

3) Encontre o valor dos parâmetros a_1, a_2 e a_3 que correspondem ao ajuste de mínimos quadrados (linearizado) da função $f(x) = a_1 x^{a_2} e^{a_3 x}$ aos dados da tabela abaixo

x_i	0,5	0,6875	0,875	1,0625	1,25	1,4375	1,625	1,8125	2
f_i	0,249	0,348	0,450	0,508	0,501	0,482	0,454	0,417	0,355

Dado que o máximo da função ajustada ocorre em $x = -\dfrac{a_2}{a_3}$, determine seu o valor máximo (arredondamento com quatro dígitos).

4) Considere os dados da tabela abaixo.

x_i	1	1,22	1,44	1,66	1,89	2,11	2,33	2,56	2,78	3
y_i	1,33	0,817	0,693	0,568	0,416	0,333	0,283	0,249	0,195	0,186

Utilize-os para realizar o ajuste linearizado da função

$$\varphi(x) = a_1 x^{a_2}.$$

O ajuste de mínimos quadrados não linear dessa mesma função é solução do sistema não linear

$$\begin{cases} \sum_{i=1}^{n} y_i\, x_i^{a_2} - \sum_{i=1}^{n} a_1\, x_i^{2a_2} & = & 0 \\ \\ \sum_{i=1}^{n} y_i\, x_i^{a_2} \ln(x_i) - \sum_{i=1}^{n} a_1\, x_i^{2a_2} \ln(x_i) & = & 0 \end{cases}.$$

Utilize o ajuste linearizado como aproximação inicial e obtenha o ajuste não linear. Determine o erro relativo cometido em cada coeficiente obtido a partir do ajuste linearizado (resposta com 4 dígitos de precisão).

5) A partir do ajuste de mínimos quadrados da função

$$\varphi(x) = a_1 \cos\frac{x}{2} + a_2 \cos x + a_3 \cos\frac{3x}{2}$$

e do ajuste de mínimos quadrados linearizado da função

$$\psi(x) = b_1 x^{b_2} \exp\left(b_3 x^2\right)$$

ao conjunto de dados abaixo, determine qual das funções resulta em um melhor ajuste sob o ponto de vista do método (ou seja, para qual função, o ajuste produz a menor soma do quadrado dos resíduos).

x_i	0,1	0,267	0,433	0,6	0,767	0,933	1,1	1,27	1,43	1,6
y_i	0,00973	0,0762	0,159	0,294	0,489	0,661	0,832	0,999	0,982	0,960

6) Utilize os dados da tabela abaixo

x	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7	7,5
ρ	1,043	2,552	7,224	13,18	18,54	22,33	22,94	20,24	14,83	10,02	6,071	2,358	1,607

para ajustar a função

$$\rho(x) = Ax^k \exp(bx + cx^2)$$

através do método dos mínimos quadrados (versão linearizada). Determine o valor das constantes com 4 dígitos.

7) Utilize os dados da tabela abaixo

x	5,5	5,72	5,94	6,17	6,39	6,61	6,83	7,06	7,28	7,5
ρ	0,1692	0,5004	1,918	4,673	6,5	7,648	7,374	5,537	2,423	1,151

para ajustar a função

$$\rho(x) = A \exp\left(-\frac{1}{2c^2}(x-b)^2\right)$$

através do método dos mínimos quadrados (versão linearizada). Determine o valor das constantes com 4 dígitos.

8) A trajetória de um satélite orbitando a Terra é descrita em coordenadas polares pela equação

$$r(\theta) = A\frac{1-\varepsilon^2}{1+\varepsilon \operatorname{sen}(\theta+\phi)},$$

onde A é o semieixo maior da órbita (em km), ε é a sua excentricidade e ϕ é uma fase. A partir do ajuste de mínimos quadrados linearizado, determine o valor aproximado (com 4 dígitos) do semieixo maior de uma órbita que passa pelos pontos indicados na tabela abaixo:

θ	- 2	- 1,5556	- 1,1111	- 0,66667	- 0,22222	0,22222	0,66667	1,1111	1,5556	2
r(km)	6758,3	6445,2	5386,0	4980,7	4728,1	4601,5	4819,2	5052,7	5430,2	6208,5

(Sugestão: a partir da identidade trigonométrica $\operatorname{sen}(\theta+\phi) = \operatorname{sen}(\theta)\cos(\phi) + \cos(\theta)\operatorname{sen}(\phi)$, é possível verificar que $\dfrac{1}{r(\theta)}$ é um combinação linear de funções de θ).

7 Integração numérica

7.1 Quadratura por interpolação

O método de quadratura por interpolação consiste em utilizar um polinômio interpolante $p(x)$ para aproximar o integrando $f(x)$ no domínio de integração $[a, b]$. Dessa forma a integral

$$\int_a^b f(x)\,dx$$

pode ser aproximada pela integral

$$\int_a^b p(x)\,dx.$$

Se o integrando $f(x)$ é conhecido em n pontos distintos $x_1, \ldots, x_n$, podemos utilizar algum dos métodos desenvolvidos para encontrar um polinômio $p(x)$ que interpole $f(x_i)$, $i = 1, \ldots, n$. Dessa forma, segundo a expressão (5.1.3):

$$\int_a^b f(x)\,dx = \int_a^b p(x)\,dx + \int_a^b dx\,\frac{f^{(n)}}{n!}\,(\zeta(x))\prod_{i=1}^{n}(x - x_i),$$

onde, a cada x, $\zeta = \zeta(x)$ é o número que torna verdadeira a equação

$$f(x) = p(x) + \frac{f^{(n)}(\zeta)}{n!}\prod_{i=1}^{n}(x - x_i).$$

De acordo com o método de interpolação de Lagrange, uma vez determinados os polinômios de Lagrange $l_i(x)$, (e a interpolação $p(x) = \sum_{i=1}^{n} f(x_i)\,l_i(x)$), a aproximação seria então dada por

$$\int_a^b p(x)\,dx = \int_a^b \sum_{i=1}^{n} f(x_i)\,l_i(x)\,dx = \sum_{i=1}^{n} f(x_i)\int_a^b l_i(x)\,dx,$$

onde a segunda igualdade se deve ao fato de que $f(x_i)$ é uma constante. A expressão anterior pode ser então reescrita na forma

$$\int_a^b p(x)\,dx = \sum_{i=1}^{n} C_i\,f(x_i),$$

onde $i = 1, \ldots, n$ e os valores $f(x_i)$ são conhecidos (fazem parte dos dados de entrada) e as constantes C_i são o resultado da integração:

$$C_i = \int_a^b l_i(x)\,dx. \tag{7.1.1}$$

A aproximação da integral de $f(x)$ é dada então por

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} C_i\, f(x_i), \tag{7.1.2}$$

onde os coeficientes C_i são dados pelas integrais (que podem ser resolvidas exatamente) (7.1.1). Essa aproximação é denominada *fórmula de quadratura*, de uma maneira geral, todas as aproximações de operações de integração numérica podem ser descritas na forma (7.1.2) – naturalmente, o coeficiente C_i vai depender do método utilizado.

Exemplo 31: Vamos aproximar a integral $\int_{-1/2}^{1/2} e^{-x^2} dx$ a partir da interpolação do integrando em três pontos: $x_1 = -1/2$, $x_2 = 0$ e $x_3 = 1/2$. Segundo o método de Lagrange, os polinômios $l_i(x)$ são:

$$l_1(x) := \frac{(x-0)(x-1/2)}{(-1/2-0)(-1/2-1/2)} = -x + 2x^2,$$

$$l_2(x) := \frac{(x+1/2)(x-1/2)}{(0+1/2)(0-1/2)} = 1 - 4x^2,$$

$$l_3(x) := \frac{(x+1/2)(x-0)}{(1/2+1/2)(1/2-0)} = x + 2x^2,$$

portanto

$$C_1 = \int_{-1/2}^{1/2} l_1(x)\,dx = \int_{-1/2}^{1/2} (-x + 2x^2)\,dx = \frac{1}{6},$$

$$C_2 = \int_{-1/2}^{1/2} l_2(x)\,dx = \int_{-1/2}^{1/2} dx\,(1 - 4x^2)\,dx = \frac{2}{3},$$

$$C_3 = \int_{-1/2}^{1/2} l_3(x)\,dx = \int_{-1/2}^{1/2} dx\,(x + 2x^2)\,dx = \frac{1}{6}.$$

Assim, a aproximação é dada pelo somatório

$$\int_{-1/2}^{1/2} e^{-x^2} dx \approx \sum_{i=1}^{3} e^{-x_i^2} C_i = \frac{1}{6}\left(e^{-\frac{1}{4}} + 4e^0 + e^{-\frac{1}{4}}\right) = 0{,}9262\ldots$$

O valor exato da integral é calculado a partir da função erro e vale

$$\int_{-1/2}^{1/2} e^{-x^2}\,dx = 0{,}9225\ldots$$

Como veremos a seguir, não será necessário construir e integrar os polinômios de Lagrange para obter a aproximação.

A chave para determinar os coeficientes é o fato de que os polinômios de Lagrange, $l_i(x)$, dependem apenas dos pontos x_i. Então, qualquer que fosse o integrando $f(x)$, uma vez fixados os pontos x_i, os polinômios de Lagrange são sempre os mesmos. Poderíamos realizar a escolha de uma função $f(x)$ dada por um polinômio, nesse caso, a interpolação é exata, ou seja, $f(x) \equiv p(x)$

e portanto

$$\int_a^b f(x)\,dx = \int_a^b p(x)\,dx = \sum_{i=1}^{n} f(x_i)\,C_i. \tag{7.1.3}$$

Em particular, vamos realizar n escolhas para a função f na forma $f_j(x) = x^j$ para $j = 0, \ldots, n-1$. Cada uma dessas escolhas para f vai originar, em vista de (7.1.3), uma equação com os n coeficientes C_i que buscamos determinar. Teremos então um sistema linear com n equações e n incógnitas[1] :

$$\begin{cases} (x_1)^0 C_1 + (x_2)^0 C_2 + \ldots + (x_n)^0 C_n &=& \int_a^b x^0\,dx &=& b-a \\ x_1 C_1 + x_2 C_2 + \ldots + x_n C_n &=& \int_a^b x\,dx &=& \frac{b^2-a^2}{2} \\ \vdots & & \vdots & & \vdots \\ (x_1)^{n-1} C_1 + (x_2)^{n-1} C_2 + \ldots + (x_n)^{n-1} C_n &=& \int_a^b x^{n-1}\,dx &=& \frac{b^n-a^n}{n} \end{cases} \tag{7.1.4}$$

Exemplo 32: Vamos utilizar os mesmos pontos do exemplo anterior, ou seja, $x_1 = -1/2$, $x_2 = 0$ e $x_3 = 1/2$. Nesse caso o sistema para os coeficientes C_i toma a seguinte forma

$$\begin{cases} C_1 + C_2 + C_3 = 1 \\ -C_1 + 0 + C_3 = 0 \\ C_1 + 0 + C_3 = \frac{1}{3} \end{cases}$$

cuja solução é $C_1 = C_3 = \dfrac{1}{6}$ e $C_2 = \dfrac{2}{3}$. Portanto a aproximação de uma integral $\int_{-1/2}^{1/2} f(x)\,dx$ é dada por

$$\int_{-1/2}^{1/2} f(x)\,dx \approx \frac{1}{6}\left(f(-1/2) + 4f(0) + f(1/2)\right).$$

Quando os pontos de interpolação $a = x_1 < x_2 < \ldots < x_n = b$ são igualmente espaçados, o método de quadratura por interpolação recebe o nome de *fórmula de Newton-Cotes*.

É interessante notar que uma vez definida a fórmula de integração em um determinado intervalo, é possível utilizar os mesmos coeficientes para aproximar a integração em outro intervalo. Basta realizar uma transformação de variável. Então se conhecemos a aproximação de uma integral $\int_a^b f(x)\,dx \approx \sum_{i=1}^n f(x_i)\,C_i = I$ e quisermos encontrar uma aproximação para $\int_c^d f(y)\,dy$, devemos realizar a mudança de variável $y = \alpha x + \beta$ que implica

$$\int_c^d f(y)\,dy = \alpha \int_{\frac{c-\beta}{\alpha}}^{\frac{d-\beta}{\alpha}} f(\alpha x + \beta)\,dx.$$

Os valores de α e β são determinados quando exigimos que os limites de integração coincidam:

[1] É possível demonstrar que se os n pontos x_i forem distintos, então o sistema possui uma única solução. Veja a referência:

Bellman, R. *Introduction to Matrix Analysis*, 2ª ed., MacGraw-Hill (1970).

$\frac{c-\beta}{\alpha} = a$ e $\frac{d-\beta}{\alpha} = b$. Ou seja,

$$\alpha = \frac{d-c}{b-a} \quad \text{e} \quad \beta = \frac{bc-ad}{b-a} \tag{7.1.5}$$

e assim,

$$\int_c^d f(y)\,dy = \alpha \int_a^b f(\alpha x + \beta)\,dx \approx \alpha \sum_{i=1}^{n} f(\alpha x_i + \beta)\,C_i$$

com α e β dados por (7.1.5)

7.2 Quadraturas newtonianas

7.2.1 Regra do trapézio

O que caracteriza as quadraturas newtonianas é o espaçamento constante entre os pontos. O caso mais simples é denominado *regra do trapézio* na qual apenas dois pontos são utilizados. De acordo com o sistema (7.1.4), a quadratura com dois pontos é dada pela fórmula

$$\int_a^b f(x)\,dx \approx C_1 f(a) + C_2 f(b),$$

onde C_1 e C_2 são solução do sistema de equações lineares

$$\begin{cases} C_1 + C_2 &= b-a \\ a\,C_1 + b\,C_2 &= \frac{b^2-a^2}{2} \end{cases}.$$

A solução do sistema é $C_1 = C_2 = \dfrac{b-a}{2}$. Se representarmos a separação entre os pontos por $h = b-a$, a regra do trapézio para a integral $\int_a^b f(x)\,dx$ assume a forma

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\left(f(a) + f(b)\right).$$

Erro de truncamento

Como já estudamos na subseção anterior, a regra do trapézio $\frac{h}{2}\left(f(a) + f(b)\right)$ para a integral de f no intervalo $[a,b]$ é o resultado da integração do polinômio $p(x)$ que interpola f nos pontos $x = a$ e $x = b$. Também estudamos no capítulo sobre interpolação que a cada x no intervalo de interpolação $[a,b]$, existe um $\xi \in (a,b)$ que depende de x (ou seja, $\xi(x)$) tal que

$$f(x) = p(x) + \frac{f^{(n)}(\xi)}{n!}\prod_{i=1}^{n}(x - x_i),$$

onde n é o número de pontos de interpolação e x_i, para $i = 1, 2, \ldots, n$ são os pontos de interpolação. Essa relação entre f e p permite estimar o erro de truncamento cometido ao aproximarmos

a integral pela regra do trapézio. Então, como

$$\frac{h}{2}\left(f(a)+f(b)\right)=\int_a^b p(x)\,dx,$$

em vista da relação entre f e p temos que

$$\begin{aligned}\int_a^b f(x)\,dx-\frac{h}{2}\left(f(a)+f(b)\right) &= \int_a^b \left(f(x)-p(x)\right)\,dx\\ &= \int_a^b \frac{f''(\xi(x))}{2}(x-a)(x-b)\,dx. \qquad (7.2.1)\end{aligned}$$

Com o objetivo de tornar explícita a dependência do termo (7.2.1) da separação entre os pontos a e b, $h=b-a$, vamos realizar a mudança de variável de integração $y=\dfrac{x-a}{h}$. Nesse caso, quando $x=a$, $y=0$ e quando $x=b$, $y=1$. Dessa forma o termo (7.2.1) pode ser reescrito como

$$\begin{aligned}\int_a^b f(x)\,dx-\frac{h}{2}\left(f(a)+f(b)\right) &= \int_0^1 \frac{f''(\xi(a+yh))}{2}h\,y\,h(y-1)\,(h\,dy)\\ &= \frac{h^3}{2}\int_0^1 f''(\xi(a+yh))\,y(y-1)\,dy \qquad (7.2.2)\end{aligned}$$

Para simplificar a última integral acima, vamos considerar ainda a seguinte forma integral do teorema do valor médio:

Teorema 7.2.1 (1º teorema do valor médio para a integração)
Se f e g são funções contínuas e g não muda de sinal no intervalo fechado $[c,d]$, então existe um ponto $\eta\in(c,d)$ tal que

$$\int_c^d f(x)g(x)\,dx=f(\eta)\int_c^d g(x)\,dx.$$

Uma vez que $y(y-1)$ não muda de sinal no intervalo $[0,1]$, o teorema garante a existência de um $\eta\in(0,1)\Rightarrow\exists\xi\in(a,b)$ tal que

$$\begin{aligned}\int_a^b f(x)\,dx-\frac{h}{2}\left(f(a)+f(b)\right) &= \frac{h^3}{2}f''(\xi)\int_0^1 y(y-1)\,dy\\ &= -\frac{h^3}{12}f''(\xi).\end{aligned}$$

Exemplo 33: Vamos estudar novamente a aproximação da integral $\int_{-1/2}^{1/2}e^{-x^2}dx$, agora porém, a partir da fórmula do trapézio para quadratura. O intervalo de integração é $\left[-\frac{1}{2},\frac{1}{2}\right]$, portanto nesse caso, $h=1$. De acordo com a fórmula do trapézio

$$\int_{-1/2}^{1/2}e^{-x^2}dx\approx\frac{1}{2}\left(e^{-1/4}+e^{-1/4}\right)=0{,}77880078\ldots$$

Quanto ao erro de truncamento na aproximação, sabemos que existe um $\zeta\in\left(-\frac{1}{2},\frac{1}{2}\right)$ tal

que

$$\int_{-1/2}^{1/2} e^{-x^2}\,dx - \frac{1}{2}\left(e^{-1/4} + e^{-1/4}\right) = -\frac{1^3}{12}\left(4\zeta^2 - 2\right)e^{-\zeta^2}.$$

A função $-\frac{1}{12}\left(4\zeta^2 - 2\right)e^{-\zeta^2}$ transforma o intervalo $\left(-\frac{1}{2}, \frac{1}{2}\right)$ no intervalo $\left(\frac{1}{12}e^{-1/4}, \frac{1}{6}\right) = (0{,}06490\ldots; 0{,}1\bar{6})$. Esse novo intervalo determina a região de possíveis valores para o erro de truncamento. De fato, a diferença entre o valor exato e a aproximação é $0{,}143761\ldots \in (0{,}06490\ldots; 0{,}1\bar{6})$.

7.2.2 Regra de Simpson

A regra de Simpson é a fórmula de quadratura de Newton com três pontos. Nesse caso, o intervalo de integração $[a, b]$ é dividido em duas partes pelo ponto intermediário $\frac{a+b}{2}$. Assim, os três pontos de interpolação x_1,x_2 e x_3 são dados por $x_1 = a$, $x_2 = a + h = \frac{a+b}{2}$ e $x_3 = a + 2h = b$, onde $h = \frac{b-a}{2}$ é a separação entre os pontos consecutivos.

A fórmula de quadratura possui a forma

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{3} C_i f(x_i),$$

onde C_i, $i = 1, 2\text{e } 3$ são solução do sistema de equações lineares

$$\begin{cases} C_1 + C_2 + C_3 &= b - a \\ aC_1 + \frac{a+b}{2}C_2 + bC_3 &= \frac{b^2-a^2}{2} \\ a^2C_1 + \left(\frac{a+b}{2}\right)^2 C_2 + b^2C_3 &= \frac{b^3-a^3}{3} \end{cases}.$$

A solução do sistema é dada por $C_1 = \frac{b-a}{6}$, $C_2 = \frac{2}{3}(b-a)$ e $C_3 = \frac{b-a}{6}$. Em termos da separação entre os pontos $h = \frac{b-a}{2}$: $C_1 = \frac{h}{3}$, $C_2 = \frac{4}{3}h$ e $C_3 = \frac{h}{3}$. Dessa forma a *regra de Simpson* é dada por

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left(f(x_1) + 4f(x_2) + f(x_3)\right). \tag{7.2.3}$$

Quanto ao erro de truncamento cometido na aproximação, o mesmo pode ser estimado de maneira análoga a que seguimos no caso da regra do trapézio: existe um $\xi \in (a, b)$ tal que

$$\int_a^b f(x)\,dx - \frac{h}{3}\left(f(x_1) + 4f(x_2) + f(x_3)\right) = -\frac{h^5}{90}f^{(4)}(\xi). \tag{7.2.4}$$

7.2.3 Regras de ordem superior

Seguindo esse programa, podemos desenvolver quadraturas com maior número de pontos, por exemplo, as quadraturas com 4 e cinco pontos possuem nome próprio. São a *regra 3/8* e a *regra de Boole:*

Regra 3/8

São utilizados 4 pontos, $x_1 = a$, $x_2 = a+h$, $x_3 = a+2h$ e $x_4 = b$, onde $h = \dfrac{b-a}{3}$. Então existe um $\xi \in (a,b)$ tal que

$$\int_a^b f(x)\,dx = \frac{3}{8}h\left(f(x_1) + 3f(x_2) + 3f(x_3) + f(x_4)\right) - \frac{3h^5}{80}f^{(4)}(\xi).$$

Regra de Boole[2]

São utilizados 5 pontos, $x_1 = a$ e $x_i = a + (i-1)h$ para $i = 2, 3, 4$ e $x_5 = b$, onde $h = \dfrac{b-a}{4}$. Existe um $\xi \in (a,b)$ tal que

$$\int_a^b f(x)\,dx = \frac{2}{45}h\left(7f(x_1) + 32f(x_2) + 12f(x_3) + 32f(x_4) + 7f(x_5)\right) - \frac{8h^7}{945}f^{(6)}(\xi).$$

No entanto devemos levar em conta que *não há garantias* de que o aumento do número de pontos implica a convergência da quadratura para o valor exato da integral[3]. Isto é um reflexo direto do fato de que as aproximações que estudamos até aqui são desenvolvidas a partir da integração de um polinômio que interpola f em pontos igualmente espaçados e, como já estudamos no capítulo sobre interpolação, existem exemplos de funções contínuas e com todas as derivadas contínuas em algum intervalo cuja interpolação polinomial com pontos igualmente espaçados não converge para f quando o número de pontos cresce (lembre-se da função de Runge $f(x) = \dfrac{1}{1+25x^2}$ no intervalo $x \in [-1,1]$).

A subseção seguinte trata de uma técnica de quadratura que garante a convergência para o valor exato da integral de f quando o número de pontos $n \to \infty$.

7.2.4 Regras compostas

Uma maneira de evitar as instabilidades relacionadas à interpolação em pontos igualmente espaçados consiste em particionar o intervalo de integração em diversos subintervalos e realizar a quadratura em cada um desses subintervalos com uma pequena quantidade de pontos. Essa ideia se assemelha à utilizada na interpolação *spline*.

Regra do trapézio composta

A regra consiste em dividir o intervalo de integração $[a,b]$ em $n-1$ sub-intervalos $[a,x_2]\bigcup [x_2,x_3]\bigcup \ldots \bigcup [x_{n-1},b] = [a,b]$, de mesma extensão $h = \dfrac{b-a}{n-1}$, isto é, $x_{k+1} - x_k = h$, para

[2] Devido a um erro tipográfico, essa regra é conhecida também como regra de Bode.

[3] Em geral, dada a forma das quadraturas estudadas até aqui, a quadratura será igual a integral de uma função f, que não seja um polinômio, apenas quando $h \to 0$, ou seja, quando o intervalo de integração for nulo.

qualquer $i = 1, 2, \ldots, n-1$; e aplicar a regra do trapézio em cada intervalo $[x_k, x_{k+1}]$. Ou seja,

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_{a=x_1}^{x_2} f(x)\,dx + \int_{x_2}^{x_3} f(x)\,dx + \ldots + \int_{x_{n-1}}^{b=x_n} f(x)\,dx \\ &\approx \frac{h}{2}\left(f(a)+f(x_2)\right) + \frac{h}{2}\left(f(x_2)+f(x_3)\right) + \ldots + \frac{h}{2}\left(f(x_{n-1})+f(b)\right) \\ &= h\left(\frac{1}{2}f(a) + f(x_2) + f(x_3) + \ldots + f(x_{n-2}) + f(x_{n-1}) + \frac{1}{2}f(b)\right),\end{aligned}$$

onde $x_1 = a$, $x_n = b$ e $x_k = a + (k-1)h$, para $k = 1, \ldots, n$.

Erro de truncamento

A cada subintervalo $[x_k, x_{k+1}]$ podemos estimar o erro de truncamento cometido na regra do trapézio: existe um $\xi_k \in (x_k, x_{k+1})$ tal que

$$\int_{x_k}^{x_{k+1}} f(x)\,dx = \frac{h}{2}\left(f(x_{k+1}) + f(x_k)\right) - \frac{h^3}{12}f''(\xi_k).$$

A união de todos os intervalos implica

$$\int_a^b f(x)\,dx = h\left(\frac{1}{2}f(a) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right) - \frac{h^3}{12}\sum_{k=1}^{n-1} f''(\xi_k). \qquad (7.2.5)$$

Se a função f'' for contínua, então existe um $\xi \in (a, b)$ tal que

$$f''(\xi) = \frac{1}{n-1}\sum_{k=1}^{n-1} f''(\xi_k).$$

Como $h = \dfrac{b-a}{n-1}$ podemos reescrever a igualdade (7.2.5) como

$$\int_a^b dx\, f(x) = h\left(\frac{1}{2}f(a) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right) - \frac{h^2}{12}(b-a)f''(\xi), \qquad (7.2.6)$$

onde $\xi \in (a, b)$. Note que nesse caso, na ausência de erros de arredondamento, a aproximação dada pela regra composta converge para a integral exata no limite $h \to 0$.

Regra de Simpson composta

De maneira totalmente análoga, podemos construir uma quadratura composta a partir da união das quadraturas realizadas nos subintervalos com três pontos igualmente espaçados. A partir de um número **ímpar** de pontos igualmente espaçados de $h = \dfrac{b-a}{n-1}$, $a = x_1 < x_2 < \ldots < x_{n-1} < x_n = b$ podemos aproximar a integral de f no intervalo $[a, b]$ pela composição das quadraturas

de Simpson nos $\dfrac{n-1}{2}$ intervalos $[a, x_3], [x_3, x_5], \ldots, [x_{n-2}, b]$:

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_{a=x_1}^{x_3} f(x)\,dx + \int_{x_3}^{x_5} f(x)\,dx + \ldots + \int_{x_{n-2}}^{b=x_n} f(x)\,dx \\
&\approx \frac{h}{3}\left(f(a) + 4f(x_2) + f(x_3)\right) + \frac{h}{3}\left(f(x_3) + 4f(x_4) + f(x_5)\right) + \ldots \\
&\quad \ldots + \frac{h}{3}\left(f(x_{n-2}) + 4f(x_{n-1}) + f(b)\right) \\
&= \frac{h}{3}\left[f(a) + 4\left(f(x_2) + f(x_4) + \ldots + f(x_{n-1})\right) + \right. \\
&\quad \left. + 2\left(f(x_3) + f(x_5) + \ldots + f(x_{n-2})\right) + f(b)\right],
\end{aligned}$$

A regra de Simpson composta pode ser representada pelo somatório

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\sum_{k=1}^{n} C_k f(x_k),$$

onde

$$C_k = \begin{cases} 1, & \text{se } k = 1 \text{ ou } k = n \\ 4, & \text{se } k \text{ for par} \\ 2, & \text{se } k \text{ for ímpar} \end{cases}.$$

A análise do erro de truncamento cometido na aproximação segue a linha já estudada na regra do trapézio composta. Cada intervalo de integração $[x_k, x_{k+2}]$ contribui com uma parcela $-\dfrac{h^5}{90} f^{(4)}(\xi_k)$, onde $\xi_k \in (x_k, x_{k+2})$ e $k = 1, 3, 5, \ldots, n-2$. Como são no total $\dfrac{n-1}{2}$ intervalos de integração, temos que

$$\int_a^b f(x)\,dx = \frac{h}{3}\sum_{k=1}^{n} C_k f(x_k) - \frac{h^5}{90}\sum_{k=1}^{\frac{n-1}{2}} f^{(4)}(\xi_k). \tag{7.2.7}$$

Se a função $f^{(4)}$ for contínua, então existe um $\xi \in (a, b)$ tal que

$$\frac{1}{\frac{n-1}{2}}\sum_{k=1}^{\frac{n-1}{2}} f^{(4)}(\xi_k) = f^{(4)}(\xi).$$

A substituição dessa última relação em (7.2.7) resulta em

$$\int_a^b f(x)\,dx = \frac{h}{3}\sum_{k=1}^{n} C_k f(x_k) - (n-1)\frac{h^5}{180} f^{(4)}(\xi),$$

como $(n-1)h = b-a$

$$\int_a^b f(x)\,dx = \frac{h}{3}\sum_{k=1}^{n} C_k f(x_k) - \frac{h^4}{180}(b-a)f^{(4)}(\xi)$$

7.2.5 Método de Romberg

O método de Romberg consiste na sucessiva aplicação da extrapolação de Richardson à quadratura do trapézio composta o que resulta em uma quadratura composta de maior exatidão.

A quadratura do trapézio composta com n pontos permite aproximar a integral de uma função f duas vezes continuamente diferenciável através da expressão

$$\int_a^b f(x)\,dx = T_n - \frac{h^2}{12}(b-a)f''(\xi),$$

onde

$$T_n = h\left(\frac{1}{2}f(a) + f(a+h) + f(a+2h) + \ldots + f(a+(n-2)h) + \frac{1}{2}f(b)\right),$$

$h = \dfrac{b-a}{n-1}$ e $\xi \in (a,b)$.

Se f for uma função de classe $\mathcal{C}^{2k+2}$ em um intervalo $[a,b]$, então de acordo com a fórmula de Euler-MacLaurin a sua integral definida nesse intervalo satisfaz a expressão

$$\int_a^b f(x)\,dx = T_n + c_2h^2 + c_4h^4 + \ldots + c_{2k}h^{2k} + c_{2k+2}\,h^{2k+2}f^{(2k+2)}(\xi), \tag{7.2.8}$$

onde T_n é a regra composta do trapézio com n pontos e espaçamento h, os coeficientes $c_2, \ldots, c_{2k}$ não dependem de h e $\xi \in (a,b)$.

Portanto, de acordo com a fórmula, uma quadratura no mesmo intervalo com $2n-1$ pontos, corresponde a um espaçamento igual a metade do original, assim

$$\int_a^b f(x)\,dx = T_{2n-1} + c_2\left(\frac{h}{2}\right)^2 + c_4\left(\frac{h}{2}\right)^4 + \ldots + c_{2k}\left(\frac{h}{2}\right)^{2k} + c_{2k+2}\left(\frac{h}{2}\right)^{2k+2} f^{(2k+2)}(\tilde{\xi}). \tag{7.2.9}$$

A extrapolação consiste em combinar as equações (7.2.9) e (7.2.8) de modo que o resultado da combinação linear cancele o termo h^2:

$$\int_a^b f(x)\,dx = \frac{4\,T_{2n-1} - T_n}{3} + d_4h^4 + \ldots + d_kh^{2k} + \tilde{\tau}_{2k+2}(h)\,h^{2k+2}.$$

A quadratura resultante, $\dfrac{4\,T_{2n-1} - T_n}{3}$ é a quadratura de Simpson composta com $2n-1$ pontos. O mesmo procedimento pode ser repetidamente iterado com o objetivo de produzir quadraturas compostas de ordem superior.

O método de Romberg propõe a seguinte abordagem. Colecionamos m quadraturas compostas pela regra do trapézio com $3, 5, 9, \ldots, 2^m + 1$ pontos. Essas quadraturas podem ser conveniente-

mente calculadas segundo a recursão:

$$T_{2^j+1} = \frac{1}{2}T_{2^{j-1}+1} + h_j \sum_{k=1}^{2^{j-1}} f\left(a + (2k-1)h_j\right), \tag{7.2.10}$$

onde $h_j = \dfrac{b-a}{2^j}$, $T_2 = h_0\left(\dfrac{1}{2}f(a) + \dfrac{1}{2}f(b)\right)$ e $j = 1, 2, \ldots m$. Como verificamos acima, de acordo com a extrapolação de Richardson, podemos encontrar a quadratura de Simpson composta com $2^j + 1$ pontos através da combinação

$$\frac{4T_{2^j+1} - T_{2^{j-1}+1}}{3}.$$

Vamos simbolizar essas novas quadraturas por $R_{j,1}$, ou seja,

$$R_{j,1} = \frac{4T_{2^j+1} - T_{2^{j-1}+1}}{3} \tag{7.2.11}$$

para $j = 1, 2, \ldots, m$. Uma nova sequência de extrapolações de Richardson cancelará os termos h^4. Denominamos essas novas quadraturas compostas por $R_{j,2}$:

$$R_{j,2} = \frac{16R_{j,1} - R_{j-1,1}}{15}.$$

Através de um processo de indução, chegamos à recorrência

$$R_{j,n} = \frac{4^n R_{j,n-1} - R_{j-1,n-1}}{4^n - 1}, \tag{7.2.12}$$

para $n = 1, 2, \ldots, j$ e onde $R_{j,0} \equiv T_{2^j+1}$. A relação de recorrência (7.2.12) é a expressão do método de Romberg. Em resumo, calculamos a quadratura do trapézio simples e as m quadraturas do trapézio compostas de acordo com a recorrência (7.2.10), em seguida, de acordo com a relação de recorrência (7.2.12), calculamos recursivamente as quadraturas $R_{j,1}$ para $j = 1, 2, \ldots, m$, $R_{j,2}$ para $j = 2, \ldots, m$, $R_{j,3}$ para $j = 3, \ldots, m$, etc. até $R_{m,m}$ que é a aproximação de ordem $O\left(h_m^{2m+2}\right)$ para a integral $\int_a^b f(x)\,dx$.

Exemplo 34: Vamos aproximar integral $\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-x^2}\,dx$ pela quadratura de Romberg $R_{4,4}$. Inicialmente será necessário calcular as quadraturas do trapézio e as compostas com $3, 5, 9$ e 17 pontos (respectivamente $T_2, T_{2+1}, T_{2^2+1}, T_{2^3+1}$ e T_{2^4+1}):

$$\begin{aligned}
T_{2^0+1} &= 0{,}7788007830714049\ldots \\
T_{2^1+1} &= 0{,}8894003915357024\ldots \\
T_{2^2+1} &= 0{,}9144067271745891\ldots \\
T_{2^3+1} &= 0{,}9205312369043574\ldots \\
T_{2^4+1} &= 0{,}9220548158903587\ldots
\end{aligned}$$

A partir dessas quadraturas podemos calcular os termos $R_{j,i}$ segundo as expressões (7.2.11)

e (7.2.12):

$$
\begin{aligned}
R_{1,1} &= 0{,}9262669276904683\ldots \\
R_{2,1} &= 0{,}9227421723875513\ldots \quad R_{2,2} = 0{,}9225071887006902\ldots \\
R_{3,1} &= 0{,}9225727401476136\ldots \quad R_{3,2} = 0{,}9225614446649510\ldots \\
R_{4,1} &= 0{,}9225626755523591\ldots \quad R_{4,2} = 0{,}9225620045793421\ldots \\
\\
R_{3,3} &= 0{,}9225623058707330\ldots \\
R_{4,3} &= 0{,}9225620134668721\ldots \quad R_{4,4} = 0{,}9225620123201903\ldots
\end{aligned}
$$

E assim,

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-x^2}\,dx \approx R_{4,4} = 0{,}9225620123201903\ldots$$

O valor exato da integral é 0,9225620128255849..., o erro está no décimo dígito.

7.3 Quadratura gaussiana

Os métodos de quadratura que envolvem a interpolação polinomial em n pontos fornecem, por construção, o valor exato da integral quando o integrando é um polinômio de grau menor ou igual $n-1$. Uma vez escolhidos os n pontos $x_i \in [a,b]$, utilizamos os n polinômios x^j, $j = 0, 1, \ldots, n-1$ para determinar os coeficientes C_i da quadratura através da solução do sistema de equações lineares

$$\sum_{i=1}^{n} C_i\,(x_i)^j = \int_a^b x^j\,dx = \frac{b^{j+1} - a^{j+1}}{j+1}.$$

A quadratura gaussiana utiliza as mesmas equações porém trata os pontos de interpolação x_i como incógnitas e inclui outras n equações relacionadas à interpolação dos polinômios x^j, $j = n, n+1, \ldots, 2n-1$. A fórmula de quadratura é determinada pela solução do sistema de $2n$ equações não lineares

$$\sum_{i=1}^{n} C_i\,(x_i)^j = \frac{b^{j+1} - a^{j+1}}{j+1} \tag{7.3.1}$$

em termos das incógnitas C_i e x_i, $i = 1, 2, \ldots, 2n$.

Como já estudamos, através de mudanças de variáveis podemos mudar o intervalo de integração. Desse modo não perdemos nenhuma generalidade ao estudar a solução do sistema não linear

(7.3.1) dado pelo limite de integração $[-1, 1]$

$$\begin{cases} C_1 + C_2 + \ldots + C_n &= \int_{-1}^{1} x^0\,dx &= 1-(-1) = 2 \\ x_1C_1 + x_2C_2 + \ldots + x_nC_n &= \int_{-1}^{1} x\,dx &= \frac{1^2-(-1)^2}{2} = 0 \\ (x_1)^2C_1 + (x_2)^2C_2 + \ldots + (x_n)^2C_n &= \int_{-1}^{1} x^2\,dx &= \frac{1^3-(-1)^3}{3} = \frac{2}{3} \\ \vdots & \vdots & \vdots \\ (x_1)^kC_1 + (x_2)^kC_2 + \ldots + (x_n)^kC_n &= \int_{-1}^{1} x^k\,dx &= \begin{cases} \frac{2}{k+1}, & \text{se } k\text{é par} \\ 0, & \text{se } k\text{é ímpar} \end{cases} \\ \vdots & \vdots & \vdots \\ (x_1)^{2n-1}C_1 + (x_2)^{2n-1}C_2 + \ldots + (x_n)^{2n-1}C_n &= \int_{-1}^{1} x^{2n-1}\,dx &= \frac{1^{2n}-(-1)^{2n}}{2n} = 0 \end{cases} \tag{7.3.2}$$

É possível demonstrar[4] que esse sistema possui apenas uma solução que satisfaça os critérios, $-1 < x_i < 1$ e $C_i > 0$. Apesar da aparente complexidade apresentada pelo sistema (7.3.2), não é difícil perceber que os pontos x_i satisfazem uma equação polinomial (basta isolar as variáveis C_i e em seguida as variáveis x_i a partir da primeira equação em (7.3.2)). Na realidade, é possível demonstrar que os pontos x_i são as raízes do polinômio de Legendre[5] de grau n, $\mathcal{P}_n$ e os coeficientes C_i são então dados pela expressão

$$C_i = \frac{2}{\left(1-x_i^2\right)\left(\mathcal{P}_n'(x_i)\right)^2}.$$

Isto permite, ao menos numericamente, construir a quadratura com um número arbitrário de pontos.

Quanto ao erro de truncamento, é possível demonstrar[6] que para funções contínuas, o método da quadratura converge para o valor exato da integral quando o número de pontos $n \to \infty$, além disso, se f for $2n$ vezes continuamente diferenciável, o erro cometido pela quadratura é dado pela expressão

$$\int_{-1}^{1} f(x)\,dx - \sum_{j=1}^{n} C_j f(x_j) = \frac{(n!)^4\,2^{2n+1}\,f^{(2n)}(\xi)}{((2n)!)^3\,(2n+1)},$$

onde $\xi \in (-1, 1)$. Em muitas situações, a expressão para o erro de truncamento na quadra-

[4] Veja a referência:
-Davis, P. ; Rabinowitz P. *Methods of Numerical Integration*, Academic Press (1975)

[5] O polinômio de Legendre de grau n, $\mathcal{P}_n(x)$ pode ser determinado através da fórmula de Rodrigues:

$$\mathcal{P}_n(x) = \frac{1}{2^n(n!)}\frac{d^n}{dx^n}\left(\left(x^2-1\right)^n\right).$$

De acordo com sua estrutura é possível determinar as raízes exatas até, pelo menos, $n = 9$.

[6] Veja a referência:
-Szidarovsky, F. ; Yakowitz, S. *Principles and Procedures of Numerical Analysis*, Plenum Press, (1978)

tura gaussiana não é de grande utilidade. Ao contrário das quadraturas compostas, conforme aumenta-se o número de pontos é necessário tomar derivadas de ordem superior, o que pode dar origem a estimativas de erro de truncamento grosseiras (a estimativa é muito maior do que o erro de truncamento típico).

As três primeiras quadraturas gaussianas no intervalo $(-1, 1)$ são dadas exatamente pelos coeficientes:

2 pontos $C_1 = C_2 = 1$ e $-x_1 = x_2 = \dfrac{1}{\sqrt{3}}$

3 pontos $C_1 = C_3 = \dfrac{5}{9}$, $C_2 = \dfrac{8}{9}$, $-x_1 = x_3 = \sqrt{\dfrac{3}{5}}$ e $x_2 = 0$.

4 pontos $C_1 = C_4 = \dfrac{1}{36}\left(18 - \sqrt{30}\right)$, $C_2 = C_3 = 1 - C_1$,
$-x_1 = x_4 = \sqrt{\dfrac{1}{35}\left(15 + 2\sqrt{30}\right)}$, $-x_2 = x_3 = \sqrt{\dfrac{1}{35}\left(15 - 2\sqrt{30}\right)}$.

5 pontos $C_1 = C_5 = \dfrac{1}{900}\left(322 - 13\sqrt{70}\right)$, $C_2 = C_4 = \dfrac{1}{900}\left(322 + 13\sqrt{70}\right)$, $C_3 = \dfrac{128}{225}$,
$-x_1 = x_5 = \sqrt{\dfrac{1}{63}\left(35 + 2\sqrt{70}\right)}$, $-x_2 = x_4 = \sqrt{\dfrac{1}{63}\left(35 - 2\sqrt{70}\right)}$ e $x_3 = 0$.

7.4 Exercícios

1) Considere a seguinte fórmula de quadratura

$$\int_0^1 f(x)\,dx \approx \frac{1}{4}f(0{,}2) + \frac{1}{2}f(0{,}5) + \frac{1}{4}f(0{,}8). \tag{7.4.1}$$

1. Qual é o grau do menor polinômio que não é integrado exatamente pela regra (7.4.1)?
2. Essa regra é uma fórmula de quadratura gerada a partir de uma interpolação polinomial? Em caso negativo construa a regra gerada a partir de uma interpolação nos pontos 0,2; 0,5 e 0,8.
3. Encontre os novos pesos para a regra (7.4.1) para o intervalo de integração: $\int_1^3 f(x)\,dx$.

2) Seja a integral

$$\int_0^1 e^{x^2}\,dx. \tag{7.4.2}$$

Encontre as estimativas inferior e superior para a quantidade mínima de subintervalos que devem ser utilizados na aproximação da integral (7.4.2) pela regra de Simpson composta de modo que a diferença entre o valor exato e a aproximação seja menor do que 10^{-6}.

3) Ao contrário do que ocorre na interpolação com pontos igualmente espaçados, a interpolação com pontos de Chebyshev não sofre de problemas de instabilidade quando aumentamos o número de pontos na interpolação. Assim, a fórmula de quadratura, desenvolvida a partir da interpolação com pontos de Chebyshev, converge para o valor exato da integral no limite em que o número de pontos tende ao infinito (em alguns casos, essa abordagem pode ser um alternativa interessante à quadratura de Gauss que requer a solução de um sistema de equações não lineares). Compare a exatidão obtida pelas fórmulas de Newton-Cotes, quadratura gaussiana e quadratura com pontos de Chebyshev ao estimar a integral, utilizando 3, 4, 5 pontos e o método de Romberg.

$$\int_0^1 \frac{1}{25x^2+1}dx.$$

Observação: a quadratura gaussiana $\int_0^1 f(x)\,dx \approx \sum_{i=1}^n C_i f(x_i)$, nesse caso é dada por:

- $n = 3$: $x_1 = 0{,}11270166537925831148207346002176$, $x_2 = 0{,}5$, $x_3 = 1 - x_1$,
  $C_1 = C_3 = \dfrac{5}{18}$ e $C_2 = \dfrac{4}{9}$.
- $n = 4$: $x_1 = 0{,}06943184420297371238802675555359 5$,
  $x_2 = 0{,}33000947820757186759866712044838$, $x_3 = 1 - x_2$, $x_4 = 1 - x1$,
  $C_1 = C_4 = 0{,}17392742256872692868653197461100$ e
  $C_2 = C_3 = 0{,}32607257743127307131346802538900$.
- $n = 5$: $x_1 = 0{,}04691007703066800360 1186560850304$,
  $x_2 = 0{,}23076534494715845448184278964990$, $x_3 = \dfrac{1}{2}$ $x_4 = 1 - x_2$, $x_5 = 1 - x_1$,
  $C_1 = C_5 = 0{,}11846344252809454375713202035996$,
  $C_2 = C_4 = 0{,}23931433524968323402064575741782$ e $C_3 = \dfrac{64}{225}$.

4) Um objeto movendo-se em um plano bidimensional localiza-se na origem no instante $t = 0$. A tabela abaixo coleta os valores de velocidade nas direções x e y a cada 5 segundos. Utilize a regra de Simpson composta e determine uma aproximação para a distância total percorrida e a posição final do objeto.

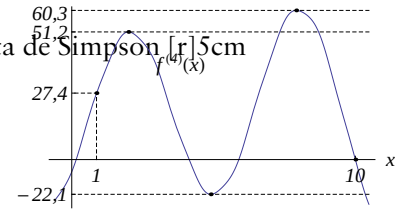
t	0	5	10	15	20	25	30	35	40	45	50
v_x	1,30	1,10	0,93	0,71	0,17	-0,62	0,15	0,92	1,10	1,40	1,20
v_y	0,71	-0,55	-0,19	0,22	0,84	1,20	-0,58	0,11	0,77	0,83	1,00

5) Construa o sistema de equações (nas variáveis C_1, C_2, C_3, α e ω) que determina a regra de quadratura por interpolação

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{3} C_i f\left(x_i\right),$$

onde $x_1 = \alpha$, $x_2 = \dfrac{1}{3}$ e $x_3 = \omega$. Supondo que f é um polinômio, qual é o maior grau para o qual essa regra é exata? Obs: não é necessário resolver o sistema.

6) A partir da regra composta de Simpson [r]5cm



$\int_1^{10} f(x)dx = S(n) - \dfrac{h^4}{180}(10-1)f^{(4)}(\eta)$, determine o menor intervalo no qual a integral exata está contida, supondo $n = 1001$, $S(n) = 1{,}771$ e que a quarta derivada de f se comporte de acordo com o gráfico ao lado.

7) Utilize a quadratura de Gauss-Legendre $\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} C_i f\left(x_i\right)$ com oito nós ($n = 8$)

i	4	3	2	1
x_i	-0,1834346424956498	-0,5255324099163290	-0,7966664774136267	-0,9602898564975362
C_i	0,3626837833783620	0,3137066458778873	0,2223810344533745	0,1012285362903763

$x_5 = -x_4, \quad x_6 = -x_3, \quad x_7 = -x_2, \quad x_8 = -x_1,$
$C_5 = C_4, \quad C_6 = C_3, \quad C_7 = C_2, \quad C_8 = C_1.$

e calcule um aproximação com seis dígitos para a integral

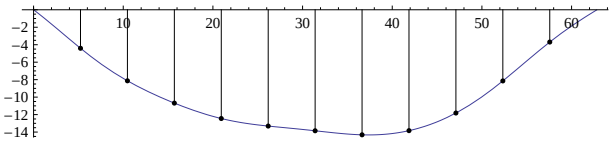
$$I = \int_1^2 \left(1 + 25x^2\right)^{-1{,}3} dx.$$

8) Utilize duas ou mais estimativas que permitam obter uma aproximação de seis dígitos para a integral definida

$$I = \int_1^5 \frac{\cos\left(\pi x\right)}{2 + \left(\operatorname{sen}\left(x\right)\right)^2} dx$$

9) A tabela abaixo contém as medidas da profundidade ao longo da seção transversal do leito de um rio com 57.3m de largura. As medidas de profundidade foram efetuadas em pontos uniformemente espaçados.

ponto	1	2	3	4	5	6	7	8	9	10	11
prof.(m)	4,4	8,1	10,7	12,4	13,3	13,9	14,3	13,8	11,8	8,1	3,7



Se considerarmos que a velocidade da correnteza é constante ao longo de toda a seção transversal do rio, teremos uma estimativa para o fluxo Φ,

$$\Phi = vA,$$

onde v é a velocidade da correnteza e A é a área da seção transversal do rio. Determine o fluxo desse rio a partir de uma aproximação para a integral via quadratura composta de Simpson, se $v = 0{,}73$m/s. (Resposta com três dígitos).

10) Os dados nas tabelas abaixa relacionam a altura, alt, ao raio, r, em uma pilha de sedimentos com simetria axial (ou seja, na forma de um sólido de revolução).

r (m)	0,00	0,75	1,50	2,25	3,00	3,75	4,50	5,25	6,00	6,75	7,50
alt(m)	25,75	25,48	24,80	23,98	23,29	22,86	22,61	22,27	21,60	20,47	18,94

r (m)	8,25	9,00	9,75	10,50	11,25	12,00	12,75	13,50	14,25	15,00
alt(m)	17,24	15,56	14,03	12,57	11,02	9,23	7,14	4,84	2,43	0,00

Utilize o método da quadratura composta de Simpson com espaçamentos $h = 0{,}75$ e $h = 1{,}5$ para determinar uma estimativa para o volume dessa pilha

$$V = 2\pi \int_0^{15} r\, alt(r) dr.$$

8 Equações Diferenciais Ordinárias

Vários modelos utilizados nas ciências naturais e exatas envolvem equações diferenciais. Essas equações descrevem a relação entre uma função, o seu argumento e algumas de suas derivadas. Como exemplo, podemos considerar o segunda lei de Newton para o movimento: o vetor momento $\mathbf{p} \in \mathbb{R}^3$ de um corpo sob ação de uma força $\mathbf{F} \in \mathbb{R}^3$ satisfaz a seguinte equação,

$$\frac{d}{dt}\mathbf{p} \equiv \mathbf{p}' = \mathbf{F}.$$

A solução da equação é uma função que depende da variável t, $\mathbf{p}(t)$, cuja derivada é diretamente proporcional à função $\mathbf{F}(t)$. Se conhecemos a dependência explícita de $\mathbf{F}(t)$ na variável t, então $\mathbf{p}(t)$ é determinada pela integral de $\mathbf{F}$. Se por outro lado $\mathbf{F}(t)$ depende também de $\mathbf{p}$, por exemplo, $\mathbf{F}(t) = \mathbf{f}(t, \mathbf{p}(t))$, então a lei de Newton, nesse caso, assume a forma da equação

$$\mathbf{p}' = \mathbf{f}(t, \mathbf{p}).$$

Essa última equação é uma *equação diferencial ordinária de* 1^a *ordem*. O termo "ordinária" indica que as derivadas presentes à equação são tomadas com respeito a uma única variável (no nosso caso, t). O termo "1ª ordem" indica que apenas $\mathbf{p}$ e sua primeira derivada estão presentes à equação.

Neste capítulo vamos estudar uma classe de equações diferenciais ordinárias denominadas *problemas de valor inicial* que caracterizam-se pela informação adicional do valor da função $\mathbf{p}(t)$ em algum $t = t_0$, ou seja, $\mathbf{p}(t_0) = \mathbf{p}_0$, em geral, essa condição é suficiente para garantir que a solução da equação

$$\begin{cases} \mathbf{p}' = \mathbf{f}(t, \mathbf{p}) \\ \\ \mathbf{p}(t_0) = \mathbf{p}_0 \end{cases}$$

é única para todo t pertencente ao intervalo em que a solução existe.

A seguir vamos estudar alguns métodos numéricos que permitem construir soluções aproximadas para as equações diferenciais ordinárias

8.1 Método da série de Taylor

Seja a equação diferencial

$$\begin{cases} y'(x) = f(x, y) \\ \\ y(x_0) = y_0 \end{cases}. \tag{8.1.1}$$

Vamos supor que a solução $y(x)$ possui uma expansão em série de Taylor em alguma vizinhança de x_0, nesse caso, a solução pode ser escrita como

$$y(x) = y(x_0) + y'(x_0)\,(x - x_0) + \frac{1}{2}y''(x_0)\,(x - x_0)^2 + \ldots + \frac{1}{n!}y^{(n)}(x_0)\,(x - x_0)^n + \ldots$$

De acordo com (8.1.1) as derivadas de y calculadas em x_0, $y^{(j)}(x_0)$, $j = 1, 2, \ldots$ podem ser calculadas pela regra da cadeia:

$$y'(x_0) = y'(x)\big|_{x=x_0} = f(x, y(x))|_{x=x_0} = f(x_0, y_0),$$

$$\begin{aligned} y''(x)\big|_{x=x_0} &= \left(\frac{\partial f}{\partial x}(x,y) + \frac{\partial f}{\partial y}(x,y)y'(x)\right)\Big|_{x=x_0} \\ &= \frac{\partial f}{\partial x}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)\, y'(x_0), \end{aligned}$$

e assim por diante para as demais derivadas. Dessa forma podemos encontrar uma aproximação para a solução $y(x)$ na vizinhança de x_0 através do truncamento da série de Taylor na ordem n e do uso das derivadas $y^{(j)}(x_0)$ calculadas através da equação (8.1.1) e da regra da cadeia.

O erro de truncamento nessa abordagem é $O\left((x - x_0)^{n+1}\right)$.

Exemplo 35: Seja a equação diferencial ordinária

$$\begin{cases} y' = x^2 + y^2 \\ \\ y(0) = 0 \end{cases}, \tag{8.1.2}$$

vamos aproximar a solução da equação pela série de Taylor em torno do ponto $x = 0$. Para tanto devemos calcular o valor das derivadas de y nesse ponto:

$$y'(0) = \left(x^2 + y(x)^2\right)\big|_{x=0} = 0^2 + y(0)^2 = 0,$$

pois $y(0) = 0$.

$$y''(0) = \left(2x + 2y(x)y'(x)\right)\big|_{x=0} = 0,$$

pois $y'(0) = 0$.

$$y^{(3)}(0) = \left(2 + 2y'(x)^2 + 2y(x)y''(x)\right)\big|_{x=0} = 2,$$

pois $y''(0) = 0$.

Até aqui temos então a aproximação

$$y(x) \approx \frac{2x^3}{3!} = \frac{x^3}{3},$$

pois $y(0) = y'(0) = y''(0) = 0$ e $y^{(3)}(0) = 2$. Vamos continuar a calcular as derivadas.

$$y^{(4)}(0) = \left(6y'(x)y''(x) + 2y(x)y^{(3)}(x)\right)\Big|_{x=0} = 0,$$

$$y^{(5)}(0) = \left(6\left(y''(x)\right)^2 + 8y'(x)y^{(3)}(x) + 2y(x)y^{(4)}(x)\right)\Big|_{x=0} = 0,$$

$$y^{(6)}(0) = \left(20y''(x)y^{(3)}(x) + 10y'(x)y^{(4)}(x) + 2y(x)y^{(5)}(x)\right)\Big|_{x=0} = 0,$$

$$y^{(7)}(0) = \left(20\left(y^{(3)}(x)\right)^2 + 30y''(x)y^{(4)}(x) + 12y'(x)y^{(5)}(x) + 2y(x)y^{(6)}(x)\right)\Big|_{x=0} = 20(2)^2 = 80.$$

Portanto os primeiros termos da série de Taylor para a solução $y(x)$ são[1]

$$y(x) = \frac{1}{3}x^3 + \frac{80}{7!}x^7 + O(x^{11}),$$

ou seja, na vizinhança de $x = 0$, a solução pode ser aproximada até a ordem $O(x^{11})$ por

$$y(x) \approx \frac{1}{3}x^3 + \frac{1}{63}x^7.$$

x_i	$\frac{1}{3}x_i^3$	$\frac{1}{3}x_i^3 + \frac{1}{63}x_i^7$	$y\,(x_i)$
0	0	0	0
0,1	0,000333	0,000333	0,000333
0,2	0,002667	0,002667	0,002667
0,3	0,009000	0,009003	0,009003
0,4	0,021333	0,021359	0,021359
0,5	0,041667	0,041791	0,041791
0,6	0,072000	0,072444	0,072448
0,7	0,114333	0,115641	0,115660
0,8	0,170667	0,173995	0,174080
0,9	0,243000	0,250592	0,250907
1,0	0,333333	0,349206	0,350232

Tabela 8.1: Comparação entre as aproximações e a solução $y(x)$ para a equação (8.1.2)

Como podemos observar pelo exemplo anterior, a aproximação se degrada conforme nos afastamos da condição inicial. O erro de truncamento é uma potência de $(x - x_0)$, portanto quanto mais afastado de x_0 estiver x, maior será o erro de truncamento cometido.

Observação 8.1.1. *Devemos lembrar que mesmo que todos os termos da série estivessem presentes, isto por si só, não garante que a série seja capaz de representar exatamente a solução $y(x)$ para qualquer x. Isto só acontece se a solução for uma função analítica em todo plano complexo. Como contra-exemplo, podemos considerar a expansão em série da função $g(x) = \dfrac{1}{1-x}$, a série*

$$1 + x + x^2 + \ldots + x^n + \ldots$$

com infinitos termos é capaz de representar $g(x)$ apenas no intervalo $|x| < 1$.

Um modo de controlar o erro de truncamento para intervalos maiores consiste em adotar uma série de pontos x_i, $i = 0, 1, 2, \ldots$ onde x_0 corresponde à condição inicial em (8.1.1), e realizar a

[1] Se continuarmos a expansão veremos que a próxima derivada não nula é $y^{(11)}(0)$.

expansão de Taylor em torno de cada ponto x_i. Dessa forma determinamos uma aproximação para a solução em uma vizinhança próxima de x_i; ao contrário do que ocorre quando aproximamos a solução pela expansão em série de Taylor em torno da origem. No entanto, para realizar as expansões em torno de x_i, devemos conhecer o valor da solução y no ponto[2] x_i, $y(x_i)$. Se $i = 0$, então $y(x_0)$ é conhecido exatamente (é a condição inicial), nos demais casos não conhecemos exatamente $y(x_i)$ mas podemos utilizar a expansão em série em torno do ponto anterior e a partir dela determinar um valor aproximado.

Vamos partir da condição inicial. No ponto x_0, a solução da EDO deve ser igual a y_0, ou seja $y(x_0) = y_0$, então a partir da equação (8.1.1) e da regra da cadeia podemos calcular o valor das derivadas $y^{(j)}(x_0)$, isto nos permite construir uma aproximação à solução $y(x)$, $\psi_0(x)$, válida na vizinhança de x_0, dada pela série de Taylor truncada no termo da n-ésima derivada:

$$\psi_0(x) = y(x_0) + y'(x_0)(x - x_0) + \ldots + \frac{1}{n!}y^{(n)}(x_0)(x - x_0)^n \approx y(x).$$

A partir dessa série encontramos uma aproximação para $y(x_1)$ dada por $y_1 = \psi_0(x_1) \approx y(x_1)$. O processo pode ser então repetido construindo uma nova série de Taylor truncada $\psi_1(x)$ em torno do ponto x_1e repetir o procedimento para encontrar a aproximação de $y(x_2)$ dada por $y_2 = \psi_1(x_2)$. Dessa forma somos capazes de construir uma iteração que aproxima o valor da solução nos pontos x_i, $i = 1, 2, \ldots$

Por simplicidade, vamos adotar a seguinte notação, $y_i^{(j)}$ é a aproximação para a $j-$ésima derivada de $y(x)$ no ponto $x = x_i$, calculada a partir de y_i, da EDO (8.1.1) e da regra da cadeia. Se os pontos x_i estiverem igualmente espaçados de h, então $x_i = x_0 + ih$. Dessa forma, podemos montar a recorrência para as aproximações da solução y nos pontos x_i.

$$y_{i+1} = y_i + y_i'h + \frac{1}{2}y_i''h^2 + \frac{1}{3!}y_i^{(3)}h^3 + \ldots + \frac{1}{n!}y_i^{(n)}h^n, \tag{8.1.3}$$

Exemplo 36: Vamos aplicar esse método à equação do exemplo anterior. Nesse caso, os pontos $x_i = x_0 + ih$, são da forma $x_i = ih$ já que no nosso caso $x_0 = 0$. De acordo com a EDO e a regra da cadeia, as derivadas são dadas por

$$y'(x) = x^2 + y(x)^2 \Rightarrow y_i' = (ih)^2 + y_i^2,$$

$$\begin{aligned} y''(x) &= 2x + 2y(x)y'(x) \\ &= 2x + 2y(x)\left(x^2 + y(x)^2\right) \\ &= 2x + 2x^2y(x) + 2y(x)^3 \\ \Rightarrow \quad y_i'' &= 2\left(ih + (ih)^2y_i + y_i^3\right), \end{aligned}$$

[2] e a partir do valor da solução em x_i e da EDO, podemos determinar o valor das demais derivadas $y'(x_i), y''(x_i), \ldots$

e

$$\begin{aligned} y^{(3)}(x) &= 2 + 2\left(y'(x)\right)^2 + 2y(x)y''(x) \\ &= 2 + 2\left(x^2 + y(x)^2\right)^2 + 2y(x)\left(2x + 2x^2y(x) + 2y(x)^3\right) \\ &= 2\left(1 + x^4 + 2xy(x) + 4x^2y(x)^2 + 3y(x)^4\right) \\ \Rightarrow\ y_i^{(3)} &= 2\left(1 + (ih)^4 + 2ihy_i + 4(ih)^2y_i^2 + 3y_i^4\right). \end{aligned}$$

Dessa forma podemos aproximar $y(x)$ até a ordem $O(h^4)$ através da relação de recorrência

$$y_{i+1} = y_i + \left((ih)^2 + y_i^2\right)h + \left(ih + (ih)^2y_i + y_i^3\right)h^2 + \frac{1}{3}\left(1 + (ih)^4 + 2ihy_i + 4(ih)^2y_i^2 + 3y_i^4\right)h^3.$$

A tabela seguinte inclui os dados obtidos pela equação de recorrência acima com $h = 0{,}1$.

x_i	$\frac{1}{3}x_i^3$	$\frac{1}{3}x_i^3 + \frac{1}{63}x_i^7$	$y_i, h = 0{,}1$	$y(x_i)$
0	0	0	0	0
0,1	0,000333	0,000333	0,000333	0,000333
0,2	0,002667	0,002667	0,002667	0,002667
0,3	0,009000	0009003	0,009003	0,009003
0,4	0,021333	0,021359	0,021357	0,021359
0,5	0,041667	0,041791	0,041784	0,041791
0,6	0,072000	0,072444	0,072433	0,072448
0,7	0,114333	0,115641	0,115630	0,115660
0,8	0,170667	0,173995	0,174025	0,174080
0,9	0,243000	0,250592	0,250810	0,250907
1,0	0,333333	0,349206	0,350064	0,350232

Tabela 8.2: Comparação entre as aproximações e a solução $y(x)$ para a equação (8.1.2)

8.2 Método de Euler

O Método de Euler consiste em construir a relação de recorrência (8.1.3) até a a ordem h, isto é, utilizamos a informação sobre $y'(x_i)$ apenas. Ou seja, $y_{i+1} = y_i + y_i'h$. Como pela equação (8.1.1), $y'(x_i) = f(x_i, y(x_i)) \Rightarrow y_i' = f(x_i, y_i)$, o Método de Euler consiste na relação de recorrência

$$y_{i+1} = y_i + hf(x_i, y_i),$$

para $i = 0, 1, 2, \ldots$, onde $y(x_0) = y_0$é a condição inicial. Uma das vantagens da fórmula de Euler consiste na desnecessidade de calcular as derivadas de y nos pontos x_i, o que pode ser uma tarefa penosa como ilustram os exemplos anteriores.

Exemplo 37: Vamos aplicar o Método de Euler para encontrar uma solução aproximada

para e EDO que temos estudado em exemplos até a agora

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

Nesse caso, os pontos $x_i = x_0 + ih$, são da forma $x_i = ih$ já que $x_0 = 0$. De acordo com o método o valor da solução y no ponto x_ié dado por y_i que satisfaz a seguinte relação de recorrência :

$$y_{i+1} = y_i + h\left((ih)^2 + y_i^2\right),$$

onde $y_0 = 0$ segundo a condição inicial.

x_i	y_i	y_i	$y(x_i)$
0	0	0	0
0,1	0	0,000328	0,000333
0,2	0,001000	0,002647	0,002667
0,3	0,005000	0,008958	0,009003
0,4	0,014003	0,021279	0,021359
0,5	0,030022	0,041664	0,041791
0,6	0,055112	0,072263	0,072448
0,7	0,091416	0,115402	0,115660
0,8	0,141252	0,173730	0,174080
0,9	0,207247	0,250438	0,250907
1,0	0,292542	0,349605	0,350232

Tabela 8.3: Comparação entre as aproximações dadas pelo Método de Euler com espaçamento $h = 0{,}1$, $h = 0{,}001$ e a solução $y(x)$ para a equação (8.1.2)

O exemplo anterior ilustra o comportamento da aproximação conforme o espaçamento entre os pontos é diminuído. Gostaríamos que quando tomado o limite $h \to 0$ o método convergisse para a solução exata (naturalmente, se não existirem erros de arredondamento). De fato, se a EDO satisfizer algumas condições gerais, isso acontece. Vamos estudar o comportamento dos erros cometidos pela aproximação quando $h \neq 0$.

Erro de truncamento

O Método de Euler consiste em truncar, a cada ponto x_i, a série de Taylor para a solução da EDO em 1ª ordem. De acordo com o Teorema de Taylor (com o termo do erro na forma de Lagrange), supondo que a solução é conhecida exatamente no ponto x_i, a expansão da solução em torno desse ponto é da forma

$$y(x) = y(x_i) + (x - x_i)\, y'(x_i) + \frac{1}{2}(x - x_i)^2\, y''(\xi),$$

onde $\xi \in (x_i, x)$. O valor da solução y calculada no ponto x_{i+1} é dado pela série anterior calculada em $x = x_{i+1}$:

$$y(x_{i+1}) = y(x_i) + hf(x_i, y(x_i)) + \frac{1}{2}h^2 y''(\xi_i),$$

onde $\xi_i \in (x_i, x_{i+1})$. Essa última equação permite que estimemos o *erro local* $\varepsilon_{i+1} := y(x_{i+1}) - y_{i+1}$ dado que y_{i+1} é calculado pelo Método de Euler:

$$y_{i+1} = y_i + hf(x_i, y_i).$$

Subtraindo as duas últimas equações temos

$$\varepsilon_{i+1} = \varepsilon_i + h(f(x_i, y(x_i)) - f(x_i, y_i)) + \frac{1}{2}h^2 y''(\xi_i). \tag{8.2.1}$$

Para continuar a análise serão necessárias algumas hipóteses sobre o comportamento da solução y e da função f.

Vamos supor que a solução y possua segunda derivada limitada no intervalo em que está definida. Assim, se a solução está definida em um intervalo $I = (x_0, x_f)$, então existe um $M < \infty$ tal que $|y''(x)| \leq M$ para todo $x \in I$. Uma outra hipótese é que a função f é Lipschitz no segundo argumento (na variável y) para todo $x \in I$, ou seja, para qualquer $x \in I$, existe um $L < \infty$, que independe de x, tal que $|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|$ (estamos supondo também que y é limitada no intervalo I).

Assumindo essas hipóteses e tomando o valor absoluto na equação (8.2.1) temos que

$$|\varepsilon_{i+1}| \leq |\varepsilon_i| + hL|y(x_i) - y_i| + \frac{1}{2}h^2 M,$$

como $|y(x_i) - y_i| = |\varepsilon_i|$temos finalmente

$$|\varepsilon_{i+1}| \leq (1 + hL)\ |\varepsilon_i| + \frac{1}{2}h^2 M. \tag{8.2.2}$$

A condição inicial é conhecida exatamente, $y(x_0) = y_0$, portanto $\varepsilon_0 = 0$ e assim, de acordo com (8.2.2)

$$|\varepsilon_1| \leq \frac{1}{2}h^2 M,$$

$$|\varepsilon_2| \leq (1 + hL)\ |\varepsilon_1| + \frac{1}{2}h^2 M \leq (1 + hL)\frac{1}{2}h^2 M + \frac{1}{2}h^2 M,$$

$$|\varepsilon_3| \leq (1 + hL)\ |\varepsilon_2| + \frac{1}{2}h^2 M \leq (1 + hL)^2\frac{1}{2}h^2 M + (1 + hL)\frac{1}{2}h^2 M + \frac{1}{2}h^2 M,$$

por indução chegamos a desigualdade para o erro acumulado na n-ésima iteração

$$|\varepsilon_n| \leq \frac{1}{2}h^2 M \sum_{j=0}^{n-1}(1 + hL)^j,$$

o somatório é uma série geométrica, $\sum_{j=0}^{n-1}(1 + hL)^j = \dfrac{(1 + hL)^n - 1}{hL}$, portanto

$$|\varepsilon_n| \leq hM\frac{(1 + hL)^n - 1}{2L},$$

considerando o fato de que $(1+x)^n \leq e^{nx}$ para quaisquer n e x positivos, temos finalmente que

$$|\varepsilon_n| \leq h\frac{M}{2L}\left(e^{Lnh}-1\right) = h\frac{M}{2L}\left(e^{L\left(x_f-x_0\right)}-1\right), \tag{8.2.3}$$

onde na última igualdade, utilizamos o fato de que $x_f = x_0 + nh$. A desigualdade (8.2.3) nos diz que a dependência do erro cometido pela aproximação de y no ponto x_f é limitado por um termo linear em h. Como M, L, x_f e x_0 são constantes finitas, então

$$\lim_{h\searrow 0}|\varepsilon_n| \leq \frac{M}{2L}\left(e^{\left(Lx_f-x_0\right)}-1\right)\lim_{h\searrow 0} h = 0.$$

Portanto, desconsiderando os erros de arredondamento, a aproximação construída pelo Método de Euler converge para a solução da EDO no limite em que o espaçamento entre os pontos se anula.

8.3 Método Runge–Kutta

A análise de erros de truncamento no Método de Euler pode ser aplicada também ao método da série de Taylor de modo semelhante ao que estudamos na seção anterior. Isto permite estabelecer as hipóteses que garantem a convergência desse método de maneira geral.

Como vimos em exemplos, uma das dificuldades na aplicação do método da série de Taylor se deve à crescente complexidade dos termos associados as j-ésimas derivadas da solução. Uma alternativa consiste em aproximar essas derivadas pelas suas derivadas numéricas.

Vamos considerar a seguinte EDO:

$$\begin{cases} y' = f(x,y) \\ \\ y\left(x_0\right) = y_0\ . \end{cases}$$

A série de Taylor da solução em torno de um ponto x_j é dada por

$$y(x) = y\left(x_j\right) + \left(x-x_j\right)y'\left(x_j\right) + \frac{1}{2}\left(x-x_j\right)^2 y''\left(x_j\right) + O\left(\left(x-x_j\right)^3\right).$$

O termo $y'\left(x_j\right)$ é calculado explicitamente através da EDO, $y'\left(x_j\right) = f\left(x_j, y\left(x_j\right)\right)$. Já o termo $y''\left(x_j\right)$ pode ser aproximado através de uma operação de diferença finita,

$$y''\left(x_j\right) \approx \frac{y'\left(x_j+\bar{h}\right)-y'\left(x_j\right)}{\bar{h}} = \frac{f\left(x_j+\bar{h}, y\left(x_j+\bar{h}\right)\right)-f\left(x_j, y\left(x_j\right)\right)}{\bar{h}}.$$

Os termos $y\left(x_j\right)$e $y\left(x_j+\bar{h}\right)$ são aproximados respectivamente por y_j e

$$y\left(x_j+\bar{h}\right) \approx y\left(x_j\right) + \bar{h}y'\left(x_j\right) \approx y_j + \bar{h}f\left(x_j, y_j\right).$$

Assim, a aproximação para segunda derivada de y em x_j assume a forma

$$y''\left(x_j\right) \approx \frac{f\left(x_j+\bar{h}, y_j+\bar{h}f\left(x_j,y_j\right)\right)-f\left(x_j,y_j\right)}{\bar{h}}.$$

Supondo então que o espaçamento $\bar{h}$ utilizado na aproximação de y_i'' é um múltiplo do espaçamento entre os pontos no método da série de Taylor, ou seja, $\bar{h} = \lambda h$, a aproximação y_{j+1} dada pela série de Taylor (8.1.3) truncada na ordem $n = 2$ é dada por

$$\begin{aligned} y_{j+1} &= y_j + hf(x_j, y_j) + \frac{1}{2}h^2 \left(\frac{f(x_j + \lambda h, y_j + \lambda h f(x_j, y_j))}{\lambda h} - \frac{f(x_j, y_j)}{\lambda h} \right) \\ &= y_j + h \left(\left(1 - \frac{1}{2\lambda} \right) f(x_j, y_j) + \frac{1}{2\lambda} f(x_j + \lambda h, y_j + \lambda h f(x_j, y_j)) \right), \end{aligned}$$

de forma simplificada temos então

$$y_{j+1} = y_j + h(b_1 k_1 + b_2 k_2), \tag{8.3.1}$$

onde $b_1 = 1 - \dfrac{1}{2\lambda}$, $b_2 = 1 - b_1$,

$$k_1 = f(x_j, y_j) \tag{8.3.2}$$

e

$$k_2 = f(x_j + \lambda h, y_j + \lambda h k_1). \tag{8.3.3}$$

as expressões dadas por (8.3.1), (8.3.2) e (8.3.3) constituem o *Método Runge-Kutta de dois estágios*. A denominação "dois estágios" se deve ao fato de que o valor y_{j+1} é calculado em dois estágios, no primeiro calculamos o termo k_1 que é utilizado no segundo estágio ao calcularmos o termo k_2.

Quando $\lambda = 1$, o método é conhecido como *Método Aperfeiçoado de Euler*, quando $\lambda = \dfrac{2}{3}$, *Método de Heun*.

Exemplo 38: Método Aperfeiçoado de Euler

Seja a EDO

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

Então o Método Aperfeiçoado de Euler consiste na aproximação y_j da solução y nos pontos $x_j = jh$, definida pela relação

$$y_{j+1} = y_j + \frac{h}{2}(k_1 + k_2),$$

onde $y_0 = 0$ (dado pela condição inicial da EDO),

$$k_1 = (x_j)^2 + (y_j)^2$$

e

$$k_2 = (x_j + h)^2 + (y_j + h\,k_1)^2.$$

A tabela seguinte ilustra o comportamento da aproximação com $h = 0{,}1$.

x_i	y_i	$y(x_i)$
0	0	0
0,1	0000500	0,000333
0,2	0,003000	0,002667
0,3	0,009503	0,009003
0,4	0,022025	0,021359
0,5	0,042621	0,041791
0,6	0,073442	0,072448
0,7	0,116817	0,115660
0,8	0,175396	0,174080
0,9	0,252374	0,250907
1,0	0,351830	0,350232

Tabela 8.4: Comparação entre a aproximação dada pelo Método Aperfeiçoado de Euler com espaçamento $h = 0{,}1$ e a solução $y(x)$ para a equação (8.1.2)

Exemplo 39: Método de Heun

Seja a EDO

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

Então o Método de Heun consiste na aproximação y_j da solução y nos pontos $x_j = jh$, definida pela relação

$$y_{j+1} = y_j + \frac{1}{4}hk_1 + \frac{3}{4}hk_2,$$

onde $y_0 = 0$ (dado pela condição inicial da EDO),

$$k_1 = (x_j)^2 + (y_j)^2$$

e

$$k_2 = \left(x_j + \frac{2}{3}h\right)^2 + \left(y_j + \frac{2}{3}hk_1\right)^2.$$

A tabela seguinte ilustra o comportamento da aproximação com $h = 0{,}1$.

x_i	y_i	$y(x_i)$
0	0	0
0,1	0,000333	0,000333
0,2	0,002667	0,002667
0,3	0,009002	0,009003
0,4	0,021355	0,021359
0,5	0,041776	0,041791
0,6	0,072411	0,072448
0,7	0,115577	0,115660
0,8	0,173913	0,174080
0,9	0,250586	0,250907
1,0	0,349640	0,350232

Tabela 8.5: Comparação entre a aproximação dada pelo Método de Heun com espaçamento $h = 0{,}1$ e a solução $y(x)$ para a equação (8.1.2)

O princípio fundamental dos métodos Runge-Kutta consiste na combinação das estimativas de derivadas com diversos espaçamentos de modo que a equação recursiva resultante possua os mesmos termos que a série de Taylor da solução até uma determinada ordem.

A generalização do métodos Runge-Kutta de n estágios é dada pela seguintes expressões:

$$y_{j+1} = y_j + h\left(\sum_{i=1}^{n} b_i k_i\right) \tag{8.3.4}$$

e os termos k_i são definidos pela seguinte equação recursiva,

$$k_1 = f(x_j, y_j)$$

e

$$k_i = f\left(x_j + c_i h, y_j + h\sum_{l=1}^{i-1} a_{i,l} k_l\right)$$

para $i = 2, 3, \ldots, n$. O ponto x_0 corresponde à condição inicial $y(x_0) = y_0$.

Os parâmetros $a_{i,l}$,c_i e b_i devem ser determinados de modo que a aproximação seja a mais exata possível.

Para tanto, vamos considerar a função $g(x_j, y_j, h)$, constituída pela soma de todos os termos $k_i(x_j, y_j, h)$ no lado direito da expressão (8.3.4):

$$y_{j+1} = y_j + h\, g(x_j, y_j, h). \tag{8.3.5}$$

Dessa forma, os parâmetros são escolhidos de modo que a série de Taylor em h para a solução em $x = x_{j+1}$, $y(x_{j+1}) = y(x_j + h)$, seja igual à serie de Taylor em h para o lado direito de (8.3.5) até um grau máximo p,

$$\sum_{m=0}^{p} \frac{1}{m!} y^{(m)}(x_j)\ h^m = y_j + h\sum_{m=0}^{p-1} \frac{1}{m!}\frac{\partial^m g}{\partial h^m}(x_j, y_j, 0)\ h^m.$$

O caso $p = 0$ implica a igualdade $y(x_j) = y_j$. Nos demais casos, a exigência de que a expressão acima seja válida para um $h \geq 0$ qualquer, implica as seguintes equações para $m = 1, 2, \ldots, p$:

$$y^{(m)}(x_j) = m \frac{d^{m-1}}{dh^{m-1}} \left(\sum_{i=1}^{n} b_i f \left(x_j + c_i h, y_j + h \sum_{l=1}^{i-1} a_{i,l} k_l \right) \right) \Bigg|_{h=0} . \tag{8.3.6}$$

Vamos estudar as situações em que temos $p = 1$ e $p = 2$. Vamos começar com um método Runge-Kutta de um único estágio, ou seja

$$y_{j+1} = y_j + \alpha_1 h \, f(x_j, y_j) .$$

Nesse caso a equação (8.3.6) com $p = 1$ implica

$$y'(x_j) = b_1 f(x_j, y_j) ,$$

como $y'(x_j) = f(x_j, y_j)$ temos que a equação é satisfeita com $b_1 = 1$. Nesse caso, o método Runge-Kutta de um único passo é idêntico ao Método de Euler. A equação para $p = 2$ não poder ser satisfeita nesse método pois $\frac{d}{dh} (b_1 f(x_j, y_j))|_{h=0} = 0$.

Vamos analisar o método Runge-Kutta de dois estágios,

$$y_{j+1} = y_j + h \left(b_1 \, f(x_j, y_j) + b_2 \, f(x_j + c_2 h, y_j + h a_{2,1} f(x_j, y_j)) \right) .$$

Nesse caso, a equação (8.3.6) com $p = 1$ implica

$$y'(x_j) = (b_1 + b_2) f(x_j, y_j) ,$$

como $y'(x_j) = f(x_j, y_j)$ temos então que

$$b_1 + b_2 = 1. \tag{8.3.7}$$

A equação (8.3.6) para $p = 2$ implica

$$y''(x_j) = 2 \left(b_2 c_2 \frac{\partial f}{\partial x}(x_j, y_j) + b_2 a_{2,1} f(x_j, y_j) \frac{\partial f}{\partial y}(x_j, y_j) \right) ,$$

como

$$y''(x_j) = \frac{\partial f}{\partial x}(x_j, y_j) + f(x_j, y_j) \frac{\partial f}{\partial y}(x_j, y_j)$$

temos então que

$$2 b_2 c_2 = 1 \tag{8.3.8}$$

e

$$2 b_2 a_{2,1} = 1. \tag{8.3.9}$$

Portanto as três equações (8.3.7), (8.3.8) e (8.3.9) devem ser satisfeitas simultaneamente. Não é difícil verificar que tanto o Método de Heun quanto o Método Aperfeiçoado de Euler as satisfazem. Na realidade, há uma infinidade de escolhas, já que são 4 incógnitas e apenas três equações.

Por outro lado é possível verificar que a equação (8.3.6) para $p = 3$ nunca pode ser satisfeita por um método Runge-Kutta de dois estágios.

Através dessa análise é possível verificar que a ordem do método Runge-Kutta é igual ao seu número de estágios até o método de 4 estágios, a partir desse número de estágios, a ordem cresce menos rapidamente do que o número de estágios, por exemplo, um método de 5 estágios possui ordem 4 como o de 4 estágios. Para obter ordem 5 são necessários 6 estágios. A ordem do método Runge-Kutta é importante pois permite obter informação sobre a taxa de convergência do método com relação ao espaçamento h. Por exemplo, sabe-se[3] que uma equação diferencial com solução no intervalo (x_0, x) aproximada nos pontos $x_1 = x_0 + h$, $x_2 = x_0 + 2h, \ldots, x = x_0 + Nh$ por $y_1, y_2, \ldots, y_N$ que satisfazem as relações de recorrência de um método Runge-Kutta de ordem p é tal que

$$\max_{0 \leq j \leq N} |y(x_0 + jh) - y_j| = O(h^p).$$

Método Runge–Kutta Clássico (4 estágios)

O Método Runge-Kutta Clássico de 4 estágios é descrito pelo seguinte conjunto de relações de recorrência:

$$y_{j+1} = y_j + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

onde

$$k_1 = f(x_j, y_j),$$

$$k_2 = f\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}hk_1\right),$$

$$k_3 = f\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}hk_2\right)$$

e

$$k_4 = f(x_j + h, y_j + hk_3).$$

A aproximação é tal que o erro absoluto no intervalo (x_0, x) é dado por

$$\max_{0 \leq l \leq N} |y(x_0 + jh) - y_j| = O\left(h^4\right)$$

onde $x = x_0 + Nh$. Ou seja, o método é de 4ª ordem.

Exemplo 40: Seja a EDO

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

Então o Método Runge-Kutta Clássico de 4 estágios consiste na aproximação y_j da solução y nos pontos $x_j = jh$, definida pela relação

$$y_{j+1} = y_j + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

[3] Veja a referência:
Henrici, P. *Discrete Variable methods in Ordinary Differential Equations*, J. Willey & Sons, Inc., (1962).

onde $y_0 = 0$ (dado pela condição inicial da EDO),

$$k_1 = (x_j)^2 + (y_j)^2,$$

$$k_2 = \left(x_j + \frac{1}{2}h\right)^2 + \left(y_j + \frac{1}{2}hk_1\right)^2,$$

$$k_3 = \left(x_j + \frac{1}{2}h\right)^2 + \left(y_j + \frac{1}{2}hk_2\right)^2$$

e

$$k_4 = (x_j + h)^2 + (y_j + hk_3)^2.$$

A tabela seguinte ilustra o comportamento da aproximação com $h = 0{,}1$.

x_i	y_i	$y(x_i)$
0	0	0
0,1	0,000333	0,000333
0,2	0,002667	0,002667
0,3	0,009003	0,009003
0,4	0,021359	0,021359
0,5	0,041791	0,041791
0,6	0,072448	0,072448
0,7	0,115660	0,115660
0,8	0,174081	0,174080
0,9	0,250908	0,250907
1,0	0,350234	0,350232

Tabela 8.6: Comparação entre a aproximação dada pelo Método Runge-Kutta Clássico com espaçamento $h = 0{,}1$ e a solução $y(x)$ para a equação (8.1.2)

8.4 Sistema de equações diferenciais de 1ª ordem

A generalização do Método de Euler e do método da série de Taylor para as solução de equações diferenciais de 1ª ordem é imediato:

Dado o problema de valor inicial

$$\begin{cases} \mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \\ \\ \mathbf{y}(x_0) = \mathbf{y}_0 \in \mathbb{R}^n \end{cases}$$

onde $x_0 \in \mathbb{R}$, $\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ e a solução é uma função com n componentes, $\mathbf{y} : \mathbb{R} \to \mathbb{R}^n$.

8.5 Métodos de múltiplos passos

Os métodos Runge-Kutta são denominados métodos de passo único pois a aproximação y_{j+1} depende explicitamente de y_j apenas. Nesta seção vamos considerar os métodos de múltiplos

passos. Um método de n passos parar aproximar a solução da EDO

$$\begin{cases} y' = f(x, y),\ x > x_0 \in \mathbb{R} \\ \\ y(x_0) = y_0 \end{cases} \tag{8.5.1}$$

possui a forma geral

$$y_{j+1} = -\sum_{l=1}^{n} \alpha_l y_{j+1-l} + h \sum_{l=0}^{n} \beta_l f(x_{j+1-l}, y_{j+1-l}), \tag{8.5.2}$$

onde $x_j = x_0 + jh$. Os coeficientes $\alpha_1, \alpha_2, \ldots, \alpha_n$ e $\beta_0, \beta_1, \ldots, \beta_n$ são os parâmetros do método.

Nos casos em que $\beta_0 \neq 0$, a relação de recorrência (8.5.2) determina um *método implícito*. Essa denominação se deve ao fato de que o valor y_{j+1} está implicitamente definido por (8.5.2) nos casos em que $\beta_0 \neq 0$. Note que nesses casos temos y_{j+1} no lado esquerdo da equação e um termo $h\beta_0 f(x_{j+1}, y_{j+1})$ no lado direito, dessa forma o valor de y_{j+1} é determinado geralmente através da solução numérica de uma equação não linear.

Uma classe muito grande de métodos de múltiplos passos é formada pelos *métodos de Adams*. Nessa classe de métodos o lado direito da EDO é substituído pela aproximação construída a a partir da interpolação polinomial em um número de pontos que corresponde ao número de passos do método. Assim , o lado direito da EDO assume a forma de um polinômio na variável independente, $p(x)$, o que permite obter a aproximação no ponto seguinte a partir do teorema fundamental do cálculo. Ou seja, no domínio $x > x_i$, a EDO no PVI (8.5.1) é aproximada por

$$y' \approx p(x), \tag{8.5.3}$$

onde $p(x)$ é construído a partir de n (que corresponde ao número de passos) aproximações anteriores para y. Em seguida, a aproximação em $x = x_i$ é obtida via integração de (8.5.3),

$$y\left(x_{i+1}\right) - y\left(x_i\right) = \int_{x_i}^{x_{i+1}} y'(x)dx \approx \int_{x_i}^{x_{i+1}} p(x)dx \qquad \underset{\text{def}}{\Longrightarrow} \qquad y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} p(x)dx.$$

A forma da aproximação evidencia uma característica comum aos métodos dessa classe: podemos notar que independente da escolha do número de passos, teremos sempre $\alpha_1 = -1$ e $\alpha_l = 0$ para todo $l \neq 1$. Outra característica importante é o fato de que fora dessa classe de métodos, os métodos de múltiplos passos não são estáveis em geral.

Os métodos de Adams subdividem-se em dois grandes grupos: os métodos explícitos ($\beta_0 = 0$), denominados métodos de *Adams-Bashforth* e os métodos implícitos ($\beta_0 \neq 0$), denominados métodos de *Adams-Moulton*.

8.5.1 Método Adams–Bashforth

O método Adams-Bashforth de n passos possui a forma

$$y_{j+1} = y_j + h \sum_{l=1}^{n} \beta_i f(x_{j+1-l}, y_{j+1-l}),$$

onde os primeiros coeficientes β_l são os da seguinte tabela

n	β_1	β_2	β_3	β_4
2	3/2	-1/2	0	0
3	23/12	-16/12	5/12	0
4	55/24	-55/24	37/24	-9/24

Tabela 8.7: coeficientes β_l no método Adams-Bashforth de n passos.

8.5.2 Método Adams–Moulton

O método Adams-Moulton de i passos possui a forma

$$y_{j+1} = y_j + h \sum_{l=0}^{n} \beta_i f(x_{j+1-l}, y_{j+1-l}),$$

onde os primeiros coeficientes β_l são os da seguinte tabela

n	β_0	β_1	β_2	β_3	β_4
1	1/2	1/2	0	0	0
2	5/12	8/12	-1/12	0	0
3	9/24	19/24	-5/24	1/24	0
4	251/720	646/720	-264/720	106/720	-19/720

Tabela 8.8: coeficientes β_l no método Adams-Moulton de $n + 1$ passos.

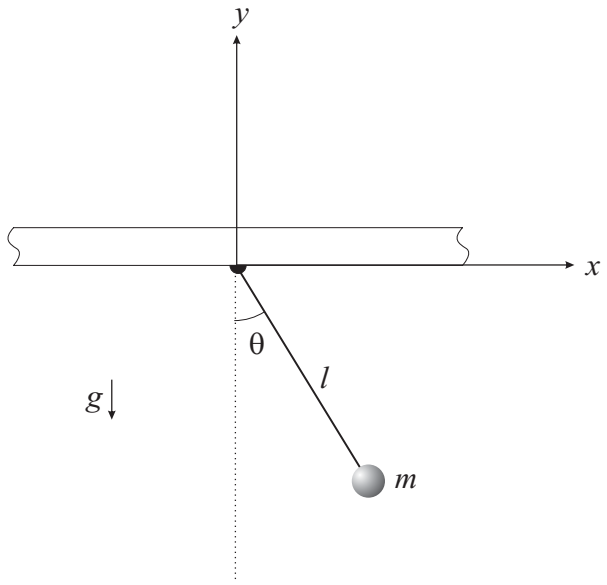
8.6 Exercícios

1) A velocidade de um corpo em queda livre pode ser descrita pela seguinte equação

$$\frac{dv}{dt} = 10 - 0{,}00343v^2.$$

A partir da equação diferencial, encontre o valor da "velocidade terminal" (valor da velocidade que implica $\frac{dv}{dt} = 0$) e através do Método Runge-Kutta Clássico de 4 estágios determine uma aproximação para o tempo que um corpo leva para alcançar 50% da velocidade terminal a partir da condição inicial $v(0) = 0$.

2) Considere o movimento do pêndulo simples em termos da função $\theta(t)$



cuja equação, de acordo com a 2ª Lei de Newton, é dada pela EDO de 2ª ordem,

$$\frac{d^2\theta}{dt^2} + \frac{g}{l}\text{sen}(\theta) = 0,$$

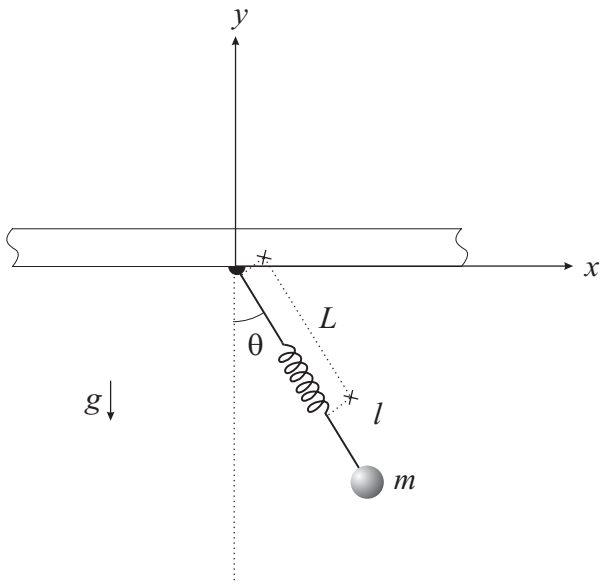
onde g é a aceleração da gravidade e l o comprimento do pêndulo. Utilize o Método Runge-Kutta Clássico de 4 estágios para determinar uma aproximação para a solução da equação do pêndulo simples nos instantes $t = 0{,}1; 0{,}2; \ldots; 1{,}0$, no caso em que $g = 10$ e $l = 1$ com condição inicial $\theta(0) = \frac{\pi}{4}, \frac{d\theta}{dt}(0) = 0$.

A energia mecânica total,

$$\frac{1}{2}ml^2\left(\frac{d\theta}{dt}\right)^2 - mgl\cos(\theta),$$

dada pela soma da energia cinética e potencial, é uma quantidade conservada ao longo do tempo. Essa propriedade é válida para a solução exata mas não para as aproximações obtidas via métodos da classe Runge-Kutta. Trabalhe com sucessivos valores para o espaçamento h e observe o comportamento da energia mecânica ao longo de t.

3) Se trocarmos o eixo rígido do pêndulo simples por um que suporte deformação elástica apenas na direção longitudinal teremos a seguinte situação



Agora, tanto o ângulo θ, quanto o comprimento l são funções do tempo. As equações que descrevem o comportamento dessas funções podem ser obtidas através das leis de Newton mas esse desenvolvimento não é o mais prático. Nesse caso, a forma mais simples de obter as equações de movimento é através do formalismo lagrangiano ou através do formalismo hamiltoniano.

As equações são dadas por

$$\begin{cases} \dfrac{d^2\theta}{dt^2} + \dfrac{2}{l}\dfrac{d\theta}{dt}\dfrac{dl}{dt} + \dfrac{g}{l}\text{sen}(\theta) & = & 0 \\ \dfrac{d^2l}{dt^2} - l\left(\dfrac{d\theta}{dt}\right)^2 + \dfrac{k}{m}(l-L) - g\cos(\theta) & = & 0 \end{cases},$$

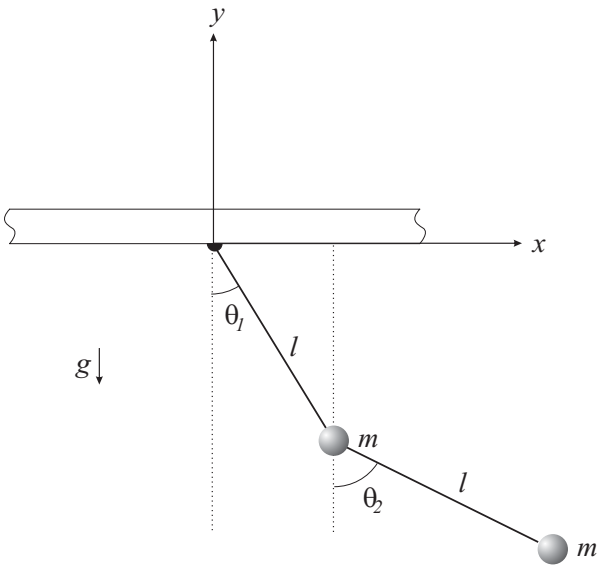
onde g é a aceleração da gravidade, k é constante de mola do eixo e L o comprimento do eixo não deformado. Utilize o Método Runge-Kutta Clássico de 4 estágios para determinar uma aproximação para a solução da equação do pêndulo simples nos instantes $t = 0{,}1; 0{,}2; \ldots; 1{,}0$, no caso em que $g = 10$, $L = 1$, $k = 2$ e $m = 0{,}2$, com condição inicial $\theta(0) = \dfrac{\pi}{4}$, $\dfrac{d\theta}{dt}(0) = 0$, $l(0) = L$, $\dfrac{dl}{dt}(0) = 0$.

Para este sistema a energia mecânica total possui a expressão

$$\frac{m}{2}\left(l^2\left(\frac{d\theta}{dt}\right)^2 + \left(\frac{dl}{dt}\right)^2\right) - mgl\cos(\theta) + \frac{1}{2}k(l-L)^2,$$

Trabalhe com sucessivos valores para o espaçamento h e observe o comportamento da energia mecânica ao longo de t.

4) Pêndulo duplo simétrico. Neste caso, o movimento é descrito pelas variáveis θ_1 e θ_2



Neste exemplo também, a forma mais simples de obter as equações de movimento é através do formalismo lagrangiano ou através do formalismo hamiltoniano.

As equações são dadas por

$$\begin{cases} \ddot{\theta_1} + \dfrac{1}{1+\text{sen}^2(\Delta\theta)}\left(\text{sen}(\Delta\theta)\left(\cos(\Delta\theta)\,\dot{\theta_1}^2 + \dot{\theta_2}^2\right) + \dfrac{g}{2l}\left(3\text{sen}(\theta_1) + \text{sen}(\Delta\theta - \theta_2)\right)\right) &= 0 \\ \ddot{\theta_2} - \dfrac{\text{sen}(\Delta\theta)}{1+\text{sen}^2(\Delta\theta)}\left(\cos(\Delta\theta)\,\dot{\theta_2}^2 + 2\dot{\theta_1}^2 + \dfrac{2g}{l}\cos(\theta_1)\right) &= 0 \end{cases}$$

onde g é a aceleração da gravidade, l o comprimento do pêndulo e $\Delta\theta = \theta_1 - \theta_2$. Utilize o Método Runge-Kutta Clássico de 4 estágios para determinar uma aproximação para a solução da equação do pêndulo simples nos instantes $t = 0{,}1; 0{,}2; \ldots; 1{,}0$, no caso em que $g = 10$ e $l = 1$ com condição inicial $\theta_1(0) = \theta_2(0) = \dfrac{\pi}{4}$, $\dfrac{d\theta_1}{dt}(0) = \dfrac{d\theta_2}{dt}(0) = 0$.

Para este sistema a energia mecânica total possui a expressão

$$\frac{ml^2}{2}\left(2\dot{\theta_1}^2 + \dot{\theta_2}^2 + 2\cos(\theta_1 - \theta_2)\,\dot{\theta_1}\dot{\theta_2}\right) - mgl\left(2\cos(\theta_1) + \cos(\theta_2)\right),$$

Trabalhe com sucessivos valores para o espaçamento h e observe o comportamento da energia mecânica ao longo de t.

5) Considere a EDO

$$m\frac{d^2x}{dt^2} = -k\,(x - l) + \frac{A}{x^2}.$$

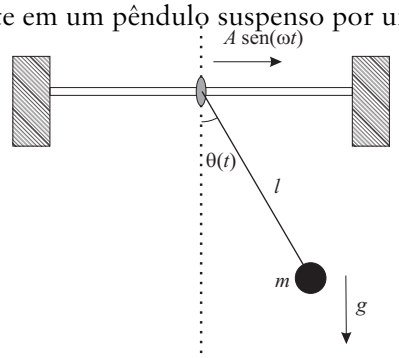
O caso $A = 0$ corresponde à EDO de um oscilador harmônico com massa m e constante de mola k (as oscilações possuem período $2\pi\sqrt{m/k}$). Determine o período das oscilações para $A = 70$, $m = 1$, $k = 10$, $l = 5$ e condições iniciais $x(0) = 5$, $x'(0) = 0$.

6) A partir do PVI

$$\begin{cases} \dot{u} = 5\,(v - u) \\ \dot{v} = (28 - w)\,u - v \\ \dot{w} = u\,v - 0{,}9w \end{cases}$$

com condição inicial $u(0) = 0$, $v(0) = 1$, $w(0) = 2$, determine a melhor aproximação com cinco dígitos para $\int_0^{10} u(t)v(t)dt$ através da quadratura composta de Simpson.

7) O seguinte sistema consiste em um pêndulo suspenso por uma haste deslizante conforme o diagrama ao lado.[r]5cm



Inicialmente o pêndulo encontra-se em repouso na posição vertical. A equação para o ângulo é

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l}\text{sen}\,(\theta) + \frac{\omega^2 A}{l}\text{sen}\,(\omega t)\cos(\theta)$$

Determine o maior ângulo alcançado no intervalo de tempo limitado por $10s$. Trabalhe com $g = 10m/s^2$, $l = 1{,}19m$, $A = 0{,}3m$ e $\omega = 2{,}5s^{-1}$.

8) Considere o seguinte P.V.I.

$$y'' + \left(e^{y'} - 1\right) + y = -3\cos(t)\,, \qquad y(0) = y'(0) = 0.$$

Obtenha uma aproximação para a solução através do Método Runge-Kutta Clássico no intervalo $t \in [0, 50]$ com espaçamento $h = 0{,}01$. A partir dos valores aproximados obtenha uma estimativa para a amplitude da oscilação no intervalo $t \in [43, 50]$ com 4 dígitos.

9) Determine a expressão explícita para a relação de recorrência dada pelo Método Aperfeiçoado de Euler no caso do sistema de EDOs

$$\begin{cases} x' = & t + x\,y \\ y' = & y + t\,x \end{cases}$$

com condição inicial $x\,(t_0) = x_0$ e $y(t_0) = y_0$.

10) O sistema de equações abaixo é um modelo simplificado para propagação de sinais elétricos por axônios.

$$\begin{cases} v' &= I + (1 - v(t))\,(-0{,}139 + v(t))\,v(t) - w(t), \qquad t > 0 \\ w' &= 0{,}008\,(v(t) - 2{,}54w(t))\,, \qquad t > 0 \\ & v(0) = 0 \qquad \text{e} \qquad w(0) = 0. \end{cases}$$

O termo I representa uma corrente externa. Através da aproximação obtida pelo Método Runge-Kutta Clássico com espaçamento $h = 0{,}01$ e excitação $I = 0{,}05$, determine se a solução evolui para a um potencial ação (v e w periódicas), ou para a situação de repouso (v e w nulas) ou ainda para a saturação (v e w constantes e não nulas).

11) Considere o seguinte P.V.I para a reentrada de um corpo na atmosfera de um planeta

$$\begin{cases} y'' &= -\dfrac{4\times 10^5}{(6400+y)^2} - 0{,}12\, y'\,|y'|\exp\left(-0{,}12y\right), \qquad t>0, y>0 \\ \\ & \qquad\qquad y(0)=100 \qquad \text{e} \qquad y'(0)=0. \end{cases}$$

A solução, $y(t)$, representa a altura no instante t dado que no instante $t=0$ o corpo se encontra em repouso a uma altura de 100 unidades de comprimento.

As equações supõem que a força de arrasto é newtoniana (quadrática na velocidade) e que a densidade de massa da atmosfera se comporta de maneira exponencial. A partir do Método Runge-Kutta Clássico, determine o instante do impacto na superfície (ou seja, para qual t^*, $y\left(t^*\right)=0$) com quatro dígitos de precisão.

12) O sistema de equações abaixo é um modelo simplificado para a evolução populacional de duas espécies interagentes

$$\begin{cases} x' &= 2\,x\,(1-y)\,, \qquad t>0 \\ y' &= 4\,y\,(x-1)\,, \qquad t>0 \\ & \quad x(0)=0{,}6 \qquad \text{e} \qquad y(0)=1{,}1 \end{cases}$$

Se $x(0)>0$ e $y(0)>0$ a solução é periódica. Determine a melhor aproximação com quatro dígitos para os valores mínimos de x e y.

13) Sejam $x(t)$ e $y(t)$ soluções do P.V.I. abaixo no domínio $(0,5]\ni t$

$$\begin{cases} x' &= 2\,x\,(1-y)\,, \qquad t>0 \\ y' &= 4\,y\,(x-1)\,, \qquad t>0 \\ & \quad x(0)=0{,}8 \qquad \text{e} \qquad y(0)=1{,}3 \end{cases}$$

Considerando as aproximações para o P.V.I. obtidas a partir do Método Runge-Kutta Clássico com 100, 200 e 400 pontos, determine a melhor aproximação que pode ser obtida para

$$\int_0^5 t\,x(t)\,y(t)\,dt$$

através da quadratura composta de Simpson.

14) A equação para a deformação de uma viga homogênea com seção transversal constante, sujeita a uma carga com distribuição uniforme q ao longo do seu comprimento e suportada simplesmente nas extremidades ($y(0)=y(L)=0$) é dada por

$$\begin{cases} \dfrac{EI}{\left(1+(y')^2\right)^{3/2}}\,y'' = -\dfrac{qLx}{2}+q\dfrac{x^2}{2}, \qquad 0<x<L, \\ \\ y(0)=y(L)=0. \end{cases}$$

Considere a situação na qual $E=2\times 10^{11}\text{N/m}^2$, $I=2{,}083\times 10^{-6}\text{m}^4$ e $q=-2500\text{N/m}$. Utilize

o Método Runge-Kutta Clássico de 4 estágios com espaçamento $h = 0{,}001$m e determine uma aproximação com quatro dígitos para o ângulo de inclinação nas extremidades da viga sob essas condições se o seu comprimento é de $L = 7{,}5$m.

9 Códigos Scilab

9.1 Eliminação Gaussiana com pivotamento parcial

```
function x=gausspp(C)

  // A matriz "C" deve ser a matriz completa de um sistema de equações li-
neares.
  // O programa possui duas partes: o escalonamento da matriz e o processo
de
  // substituição.

  // Variáveis auxiliares
  //
  // n -> número de linhas da matriz completa.
  // m -> número de colunas da matriz completa.
  // i -> indexador de linha.
  // j -> indexador de coluna.
  //
  // pivot -> guarda o elemento pivô.
  // pivot_line -> guarda o índice de linha de um elemento pivô.
  // per_line -> variável booliana que indica a necessidade de trocar uma
linha.
  // aux_line -> guarda uma das linhas a ser trocada.
  // fator_num -> fator utilizado na eliminação dos elementos abaixo do pivô.
  // fator_d -> fator utilizado na eliminação dos elementos abaixo do pivô.

  n=size(C,1) // Número de linhas de C.
  m=size(C,2) // Número de colunas de C.

  //
  // Escalonamento da matriz de coeficientes
  //
  for j=1:(n-1)
    pivot=C(j,j) // Inicialmente consideramos o elemento da diagonal como
o pivô.
    per_line=%F // Inicialização da variável booliana.
    // verificação da necessidade de troca de linhas
    for i=j+1:n
      per_line=abs(C(i,j))>2*abs(pivot) // Um outro elemento na coluna é maior
do
      if per_line then // que o candidato a pivô anterior.
        pivot_line=i // Guardamos o índice da linha para a troca.
        pivot=C(i,j) // Atualizamos o pivô.
```

```
      end
    end
    // "Pivoteamento" parcial
    if per_line then // Se houver necessidade de troca, a variável
      aux_line=C(j,:) // auxiliar guarda a linha mais "alta".
      C(j,:)=C(pivot_line,:) // Troca da linha.
      C(pivot_line,:)=aux_line // Troca da linha.
    end
    // Eliminação dos elementos na coluna, abaixo do pivô.
    fator_den=pivot
    for i=j+1:n
      fator_num=C(i,j)
      C(i,j)=0
      C(i,j+1:m)=-(fator_num/fator_den)*C(j,j+1:m) + C(i,j+1:m)
    end
  end

  //
  // Substituição
  //
  x(n,1:m-n)=C(n,n+1:m)/C(n,n)
  for i=n-1:-1:1
    x(i,1:m-n)=(C(i,n+1:m)-C(i,i+1:n)*x(i+1:n,1:m-n))/C(i,i)
  end
  endfunction
```

9.2 Método de Jacobi

```
function [x]=jacobi_solv(A,norma,tol,Nmax)

  // Calcula a solução aproximada de um sistema de equações lineares
  // via método de Jacobi.
  // O sistema a*x=b deve estar na forma de uma matriz completa A:=[a b].
  // A matriz de coeficientes "a" é decomposta como a=B-C onde "B" é a ma-
triz
  // diagonal formada pela diagonal principal de "a".

  // Variáveis de entrada
  // A -> matriz completa de um sistema de equações lineares.
  // norma -> norma matricial utilizada nas estimativas de convergência.
  // tol -> tolerância na diferença entre duas aproximações sucessivas.
  // Nmax -> número máximo de iteradas.

  // Variável de saída
  // x -> solução aproximada do sistema.

  // Variáveis auxiliares
  // n -> número de linhas no sistema.
  // m -> número de colunas no sistema.
  // InvB -> inversa de B.
  // C -> matriz C=B-a.
```

```
  // InvB_b -> produto InvB*b.
  // InvB_C -> produto InvB*C.
  // x0 -> solução aproximada do sistema na iterada anterior.
  // segue -> variável booliana. Controla a execução das iteradas.
  // contador -> contador do número de iteradas.

  // Construção das matrizes utilizadas na iteração
  n=size(A,1)
  m=size(A,2)

  invB=zeros(n,n)
  for i=1:n
    invB(i,i)=1/A(i,i)
  end

  C=-A(:,1:n)
  for i=1:n
    C(i,i)=0
  end

  invB_b=zeros(n,m-n)
  for i=1:n
    invB_b(i,1:m-n)=invB(i,i)*A(i,n+1:m)
  end

  invB_C=zeros(n,n)
  for i=1:n
    invB_C(i,1:n)=invB(i,i)*C(i,1:n)
  end

  // Aproximação inicial e inicialização das demais variáveis.
  x0=zeros(n,m-n)
  segue=%T
  contador=0

  // Iteração
  while segue
    contador=contador+1
    x=invB_b+invB_C*x0
    segue=~(norm(x-x0,norma)<=tol|contador==Nmax)
    x0=x
  end

  if contador>=Nmax then
    warning('Não houve convergência.')
  end
  endfunction
```

9.3 Método Gauss-Seidel

```
function [x]=gseidel_solv(A,norma,tol,Nmax)

  // Calcula a solução aproximada de um sistema de equações lineares
```

```
  // via método Gauss-Seidl.
  // O sistema a*x=b deve estar na forma de uma matriz completa A:=[a b].
  // A matriz de coeficientes "a" é decomposta como a=B-C onde "B" é a ma-
triz
  // diagonal formada pela diagonal principal de "a".

  // Variáveis de entrada
  // A -> matriz completa de um sistema de equações lineares.
  // norma -> norma matricial utilizada nas estimativas de convergência.
  // tol -> tolerância na diferença entre duas aproximações sucessivas.
  // Nmax -> número máximo de iteradas.

  // Variável de saída
  // x -> solução aproximada do sistema.

  // Variáveis auxiliares
  // n -> número de linhas no sistema.
  // m -> número de colunas no sistema.
  // InvB -> inversa de B.
  // C -> matriz C=B-a.
  // InvB_b -> produto InvB*b.
  // InvB_C -> produto InvB*C.
  // x0 -> solução aproximada do sistema na iterada anterior.
  // segue -> variável booliana. Controla a execução das iteradas.
  // contador -> contador do número de iteradas.

  // Construção das matrizes utilizadas na iteração
  n=size(A,1)
  m=size(A,2)

  invB=zeros(n,n)
  for i=1:n
    invB(i,i)=1/A(i,i)
  end

  C=-A(:,1:n)
  for i=1:n
    C(i,i)=0
  end

  invB_b=zeros(n,m-n)
  for i=1:n
    invB_b(i,1:m-n)=invB(i,i)*A(i,n+1:m)
  end

  invB_C=zeros(n,n)
  for i=1:n
    invB_C(i,1:n)=invB(i,i)*C(i,1:n)
  end

  // Aproximação inicial e inicialização das demais variáveis.
  x0=zeros(n,m-n)
  segue=%T
```

```
  contador=0

  // Iteração
  while segue
    contador=contador+1
    for i=1:n
      x(i)=invB_b(i)+invB_C(i,:)*x0
    end
    segue=~(norm(x-x0,norma)<=tol|contador==Nmax)
    x0=x
  end

  if contador>=Nmax then
    warning('Não houve convergência.')
  end
  endfunction
```

9.4 Método da Bissecção

```
function [z,fz,niter]=fsolve_b(x0,xl,%fun,tol,Nmax) // Método da bissec-
ção

  // Variáveis de entrada
  // x0 -> extremidade inferior.
  // xl -> extremidade superior.
  // %fun -> função.
  // tol -> tolerância na diferença relativa entre duas aproximações suces-
sivas.
  // Nmax -> número máximo de iteradas.

  // Variáveis de saída
  // z -> solução aroximada.
  // fz -> valor de %fun em z.
  // niter -> número de iteradas.

  // Variáveis auxiliares
  // f0 -> valor da função e em x=x0.
  // fl -> valor da função e em x=xl.
  // contador -> conta o número da vezes que o laço principal é executado.
  // segue -> variável booliana que controla a parada.
  // xm -> ponto intermediário.
  // fm -> valor da função f em x=xm.

  // Inicialização das variáveis
  f0=%fun(x0)
  fl=%fun(xl)
  contador=0
  segue=%T
  // Caso não tenha sido definida na chamada da função, a
  // tolerância recebe o valor 1e-10.
  if ~isdef('tol','local') then
```

```
    tol=1e-10
  end
  // Caso não tenha sido definida na chamada da função, a
  // Nmax recebe o valor 100.
  if ~isdef('Nmax','local') then
    Nmax=100
  end

  // Checagem inicial
  if f0*f1>0 then
    warning('O intervalo inicial pode não conter solução.')
    segue=%F
  end

  // Laço principal
  while segue
    xm=0.5*(x0+x1)
    fm=%fun(xm)
    // Escolha das novas extremidades
    if fm*f0<0 then
      x1=xm
      f1=fm
    else
      x0=xm
      f0=fm
    end
    contador=contador+1
    // Teste para o prosseguimento do laço
    segue=~(fm==0|abs(x1-x0)< tol*abs(xm)|contador==Nmax)
    // “Descomente” a linha abaixo se quiseres a sequência de iterações no
console
    //mprintf('Iteração %d \t aproximação: %.18e\n',contador,xm)
  end

  // Saída de dados

  // mensagem de aviso
  if contador==Nmax then
    warning(msprintf('A exatidão não foi obtida em %d iteracoes.',Nmax))
  end

  z=xm
  fz=fm
  niter=contador
  endfunction
```

9.5 Método Newton–Raphson

```
function [z,fz,niter]=fsolve_nr(x0,%fun,%dfun,tol,Nmax) // Método Newton-
Raphson
```

```
// variáveis de entrada
// x0 -> aproximação inicial.
// %fun -> função.
// %dfun -> derivada (ou matriz jacobiana) da função. Parâmetro opcional.
// tol -> tolerância na diferença relativa entre duas aproximações
// sucessivas.
// Nmax -> limite superior para o número de iteradas.

// variáveis de saída
// z -> solução aproximada.
// fz -> valor de %fun em z.
// niter -> número de iteradas.

// variáveis auxiliares
// dim -> dimensão do sistema de equações.
// contador -> guarda o número de iteradas realizadas.
// dernum -> variável booliana. É igual a %T se %dfun estiver presente.
// df0 -> valor da derivada (ou determinante da jacobiana) em x0.
// segue -> variável booliana. Controla o fluxo de execução das iteradas.
// xl -> aproximação na iterada seguinte.
// cr_conv -> variável booliana. É igual a %T se ocorrer convergência.
// abs_difl -> diferença absoluta entre duas aproximações sucessivas.
// abs_df0 -> diferença absoluta entre duas aproximações sucessivas na
// iterada anterior.

// inicialização das variáveis
dim=size(x0,1)
contador=1
dernum=~isdef('%dfun','local')

// Caso não tenha sido definida na chamada da função, a
// tolerância recebe o valor 1e-10.
if ~isdef('tol','local') then
  tol=1e-10
end

// Caso não tenha sido definida na chamada da função, a
// Nmax recebe o valor 100.
if ~isdef('Nmax','local') then
  Nmax=100
end

// A primeira iterada é calculada fora do laço principal.
// Isto é necessário para avaliar a evolução do resíduo
// já na 1ª iteração.
if dernum then
  warning('As derivadas serão estimadas numericamente.')
  warning('Existe a possibilidade de ocorrerem erros significativos.')
  df0=numderivative(%fun,x0)
else
  df0=%dfun(x0)
```

```
end
segue=det(df0)<>0
if segue then
  xl=x0-df0\%fun(x0)
else
  warning('A aproximação não pôde ser obtida.')
  warning('Verifique as aproximações iniciais ou o condicionamento.')
end
// "Descomente" a linha abaixo se quiseres a sequência de iterações
// no console:
//mprintf('Iteração %d \t aproximação: %.18e\n',contador,xl)
abs_dif0=abs(x0-xl)
x0=xl

// Laço principal
if dernum then
  df0=numderivative(%fun,x0)
  segue=det(df0)<>0
  while segue
    contador=contador+1
    xl=x0-df0\%fun(x0)
    abs_difl=norm(x0-xl,1)
    // "Descomente" a linha abaixo se quiseres a sequência de iterações
    // no console:
    //mprintf('Iteração %d \t aproximação: %.18e\n',contador,xl)
    x0=xl
    df0=numderivative(%fun,x0)
    cr_conv=abs_difl<=tol*norm(xl,1)|(abs_difl>=abs_dif0 & contador>=4)
    segue=~(cr_conv|contador>=Nmax|det(df0)<>0)
    abs_dif0=abs_difl
  end
else
  df0=%dfun(x0)
  segue=det(df0)<>0
  while segue
    contador=contador+1
    xl=x0-df0\%fun(x0)
    abs_difl=norm(x0-xl,1)
    // "Descomente" a linha abaixo se quiseres a sequência de iterações
    // no console:
    //mprintf('Iteração %d \t aproximação: %.18e\n',contador,xl)
    x0=xl
    df0=%dfun(x0)
    cr_conv=abs_difl<=tol*norm(xl,1)|(abs_difl==abs_dif0 & contador>=4)
    segue=~(cr_conv|contador>=Nmax|det(df0)==0)
    abs_dif0=abs_difl
  end
end
```

```
  // Saída de dados

  // Mensagem de aviso
  if contador==Nmax then
    warning(msprintf('\n A exatidão não foi alcançada em %d iterações.',Nmax))
  end

  z=xl
  fz=%fun(z)
  niter=contador
  endfunction
```

9.6 Método da Secante

```
function [z,fz,niter]=fsolve_s(x0,xl,%fun,tol,Nmax)// Método da Secante

  // variáveis de entrada
  // x0 -> primeira aproximação.
  // xl -> segunda aproximação.
  // %fun -> função
  // tol -> tolerância na diferença relativa entre
  // duas aproximações sucessivas.
  // Nmax -> limite superior para o número de iteradas.

  // variáveis de saída
  // z -> solução aproximada.
  // fz -> valor de %fun em z.
  // niter -> número de iteradas.

  // variáveis auxiliares
  // f0 -> valor da função f em x0.
  // fl -> valor da função f em xl.
  // contador -> guarda o número de iteradas realizadas.
  // segue -> variável booliana. Controla o fluxo de execução das iteradas.
  // x2 -> aproximação na iterada seguinte.
  // cr_conv -> variável booliana. É igual a %T se ocorrer convergência.
  // abs_difl -> diferença absoluta entre duas aproximações.
  // abs_dif0 -> diferença absoluta na iteração anterior.

  // inicialização das variáveis
  f0=%fun(x0)
  fl=%fun(xl)

  // Caso não tenha sido definida na chamada da função, a
  // tolerância recebe o valor 1e-10.
  if ~isdef('tol','local') then
    tol=1e-10
  end

  // Caso não tenha sido definida na chamada da função, a
  // Nmax recebe o valor 100.
  if ~isdef('Nmax','local') then
```

```
  Nmax=100
end
contador=1
if x0<>x1 & f0==f1 then
  warning('A aproximação não pôde ser obtida.')
  warning('Verifique as aproximações iniciais ou o condicionamento.')
  segue=%F
else
  x2=x1-((x1-x0)/(f1-f0))*f1
  if x2==x1 | x2==x0 then
    segue=%F
  else
    segue=%T
    abs_dif0=abs(x2-x1)
    x0=x1
    f0=f1
    x1=x2
    f1=%fun(x1)
    // "Descomente" a linha abaixo para exibir a sequência de iterações.
    //mprintf('Iteração %d \t aproximação: %.18e\n',contador,x1)
  end
end

// Laço principal
while segue
  if abs(f1)>abs(f0) then
    x2=x1-((x1-x0)/(1-f0/f1))
  else
    x2=x1-(x0-x1)/(1-f1/f0)*(f1/f0)
  end
  abs_dif1=abs(x2-x1)
  x0=x1
  f0=f1
  x1=x2
  f1=%fun(x1)
  contador=contador+1
  // "Descomente" a linha abaixo para exibir a sequência de iterações.
  //mprintf('Iteração %d \t aproximação: %.18e\n',contador,x1);
  cr_conv=abs_dif1<=tol*abs(x1)|(abs_dif1>=abs_dif0 & contador>=6)
  segue=~(cr_conv|f1==0|f0==f1|contador==Nmax)
  abs_dif0=abs_dif1
end

// Saída de dados

// mensagens de aviso
if contador==Nmax then
  warning(msprintf('\nA exatidão não foi alcançada em %d iteracoes.\n',Nmax))
end
```

```
z=xl
fz=fl
niter=contador
endfunction
```

10 Respostas de alguns exercícios

10.1 Capítulo 1

1)

1. $\dfrac{749}{8} = 93{,}625.$

2. $\dfrac{63}{8} = 7{,}875.$

3. $\dfrac{22015}{128} = 171{,}9921875.$

2) $0{,}0001110000\ldots_2 = 0{,}0022220120\ldots_3 = 0{,}1C28F5C28F\ldots_{16}$.

3) São necessários 13bits. Um registro de 13 bits é capaz de representar 8191 inteiros (com sinal):

$$-4095, -4095, \ldots, 0, \ldots, 4095, 4095.$$

4) Base 2 e base 4. $(111{,}0101)_2$ e $(13{,}11)_4$.

5) Serão $1 + 9 \cdot 10^4 \cdot 96 = 1 + 864 \cdot 10^4$ soluções se o arredondamento for por truncamento e $2 + 4 \cdot 10^4 + 9 \cdot 10^4 \cdot 95 = 2 + 859 \cdot 10^4$ soluções se o truncamento for par. O resultado é obtido levando-se em conta todas as combinações operações da forma

$$\begin{array}{rl} & 1{,}0000 \times 10^0 \\ \oplus & 0{,}0000 d_1 d_2 \ldots d_5 \times 10^e \end{array}$$

com $e \leq 1$.

7) No primeiro caso, a sequência de operações é

$$\begin{aligned} (x \otimes x) \ominus 1 &= (1{,}00010 \times 10^0 \otimes 1{,}00010 \times 10^0) \ominus 1{,}00000 \times 10^0 \\ &= 1{,}00020 \times 10^0 \ominus 1{,}00000 \times 10^0 \\ &= 2{,}00000 \times 10^{-4}. \end{aligned}$$

No segundo caso,

$$\begin{aligned}(x \oplus 1) \otimes (x \ominus 1) &= (2{,}00010 \times 10^0) \otimes (1{,}00000 \times 10^{-4}) \\ &= 2{,}00010 \times 10^{-4}.\end{aligned}$$

No terceiro caso,

$$\begin{aligned}(x \otimes (x \ominus 1)) \oplus (x \ominus 1) &= (1{,}00010 \times 10^0 \otimes 1{,}00000 \times 10^{-4}) \oplus 1{,}00000 \times 10^{-4} \\ &= 1{,}00010 \times 10^{-4} \oplus 1{,}00000 \times 10^{-4} \\ &= 2{,}00010 \times 10^{-4}.\end{aligned}$$

Portanto, para $x = 1{,}00010 \times 10^0$, os dois últimos casos são os mais exatos.

8) O registro hexadecimal $ABCD_h$ corresponde ao binário guarda o registro de ponto flutuante IEEE754 de 16 bits: 1010101111001101. Por sua vez, esse registro, 1010101111001101 corresponde ao ponto flutuante

$$(-1)^1(1{,}1111001101)_2 \times 2^{(01010)_2 - 15}$$

cujo valor em base decimal é

$$-6{,}0943603515625 \times 10^{-2}.$$

O ponto flutuante que representa 1 é dado por

$$(1{,}0000000000)_2 \times 2^0$$

como estamos utilizando o arredondamento par, o menor ponto flutuante (não normalizado) a ser adicionado é

$$(0{,}00000000010000000001)_2 \times 2^0$$

que corresponde a $2^{-11} + 2^{-21}$.

9) Os dois números são $(0{,}2121)_4 \times 4^{-1}$ e $(0{,}2111)_4 \times 4^0$. Antes de realizar o arredondamento a soma é $(0{,}02121)_4 + (0{,}2111)_4 = (0{,}23231)_4$ que arredondada para 4 dígitos de significando corresponde a $(0{,}2323)_4 = 0{,}73046875$.

11) $\dfrac{\Delta f_0}{f_0} \approx 0{,}075 = 7{,}5\%$.

12) A expressão $2 + x - \sqrt{x^2 + 4}$ estará sujeita a cancelamento catastrófico, quando x assumir valores muito pequenos[1]. Nesse caso o argumento da raiz será muito próximo de 4 e o resultado da expressão será igual a 0.

[1] Também há cancelamento nessa expressão quando x for um real positivo muito grande. A análise é semelhante.

Vamos então passar o termo 4 para fora da raiz,

$$2 + x - 2\sqrt{\frac{x^2}{4} + 1}.$$

Nessa nova forma, podemos verificar que o argumento da raiz possui o comportamento $\sqrt{1+z}$ com z pequeno se x for grande. Isso motiva a expansão em série de potências da raiz:

$$\sqrt{1+z} = 1 + \frac{1}{2}z - \frac{1}{8}z^2 \ldots$$

Assim, a expressão original é reescrita na forma da expansão

$$2 + x - 2\left(1 + \frac{x^2}{8} - \frac{x^4}{128} + \ldots\right)$$

a partir da qual construímos a aproximação, válida para $x \ll 1$,

$$2 + x - \sqrt{x^2 + 4} \approx x - \frac{x^2}{4} + \frac{x^4}{64}.$$

No caso do sistema $F(10, 10, -20, 20)$ e para os valores de x pedidos, basta utilizar os dois primeiros termos da aproximação:

$$x - \frac{x^2}{4}.$$

As aproximações serão dadas então por

$$\begin{aligned} &9{,}999997500 \times 10^{-7} \quad , \\ &9{,}999999750 \times 10^{-8} \quad , \\ &9{,}999999975 \times 10^{-9} \quad . \end{aligned}$$

A título de curiosidade, os cem primeiros dígitos exatos da expressão calculada com $x = 1 \times 10^{-7}$ são

$$\begin{aligned} &9{,}99999750000000000015624999999998046875000000030517578124994659423828126001358032226365804672241250890 \ldots \times 10^{-8} \quad . \end{aligned}$$

13) O termo $\cos y$ é limitado (pois a função cosseno admite valores no intervalo $[-1, 1]$). Por outro lado, $e^{\frac{x^2}{2}}$ e e^{x^2} pode assumir valores arbitrariamente grandes e é sempre maior ou igual à unidade. Além disso, $\sqrt{e^{x^2}} = e^{\frac{x^2}{2}}$. Assim,

$$\begin{aligned} e^{\frac{x^2}{2}} - \sqrt{e^{x^2} + \cos y} &= e^{\frac{x^2}{2}} - e^{\frac{x^2}{2}}\sqrt{1 + e^{-x^2}\cos y} \\ &= e^{\frac{x^2}{2}} - e^{\frac{x^2}{2}}\left(1 + \frac{1}{2}e^{-x^2}\cos y - \frac{1}{8}e^{-2x^2}\cos^2 y + \ldots\right), \end{aligned}$$

onde, na última linha, utilizamos uma expansão em série de potências na qual $e^{-x^2}\cos y$ é considerado um número pequeno. Dessa forma, levando em conta apenas os primeiros termos da expansão, chegamos à aproximação

$$e^{\frac{x^2}{2}} - \sqrt{e^{x^2} + \cos y} \approx -\frac{1}{2}e^{-\frac{x^2}{2}}\cos y + \frac{1}{8}e^{-\frac{3x^2}{2}}\cos^2 y.$$

14) A representação do número 20,25 em base três é $202{,}\overline{02}_3$. Ou seja,

$$2{,}0202020202020\ldots_3 \times 3^2,$$

onde oito primeiros dígitos da representação estão assinalados em vermelho. Se dispormos os demais dígitos após uma virgula teremos

$$0{,}20202\ldots_3 > \frac{1}{2}.$$

Portanto, de acordo com o arredondamento par, a representação em ponto flutuante com 8 dígitos de significando para o número 20,25 é

$$2{,}0202021_3 \times 3^2.$$

15) De acordo com a fórmula para propagação de erros, se os erros em x e y forem descorrelacionados,

$$\begin{aligned}
\Delta f &\approx |\text{sen}(x)\cos(y)|\,\Delta x + |\cos(x)\text{sen}(y)|\,\Delta y \\
&= \left|\frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2}\right|\Delta x + \left|\frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2}\right|\Delta y \\
&= 10^{-6}.
\end{aligned}$$

Portanto,

$$\frac{\Delta f}{f}\left(\frac{\pi}{4}, \frac{\pi}{4}\right) \approx \frac{10^{-6}}{2{,}5} = 0{,}4 \times 10^{-6}.$$

16) O número 6,125 possui a representação binária $110{,}001_2$. Em notação normalizada com sete dígitos (já que seis serão armazenados no registro)

$$1{,}100010_2 \times 2^2,$$

os dígitos que serão armazenados estão grafados em vermelho. O expoente vale 2 o deslocamento de três unidades resulta no número 5 cuja representação binária é 101_2. Finalmente o sinal do número é positivo e portanto o dígito que o representa é o 0. Assim, o registro de 10 bits é

$$0101100010$$

- 1 dígito (em azul) representa o sinal.
- 2º ao 4º dígito (em verde) representam o expoente deslocado de três unidades.
- 5º ao 10º digito (em vermelho) representam os dígitos do significando com exceção do mais significativo.

10.2 Capítulo 2

3) $\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|_1}{\|\mathbf{x}\|_1} \approx \left\|\mathrm{A}^{-1}\right\|_1 \|\mathrm{A}\|_1 \frac{\left\|\mathbf{b}-\tilde{\mathbf{b}}\right\|_1}{\|\mathbf{b}\|_1}$. Como $\left\|\mathrm{A}^{-1}\right\|_1 \|\mathrm{A}\|_1 \approx 144$ e $\left\|\mathbf{b}-\tilde{\mathbf{b}}\right\|_1 \leq \|(0{,}001; 0{,}001)\|_1$ temos que $\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|_1}{\|\mathbf{x}\|_1} \approx 0{,}24$.

5) A convergência é muito lenta. Por exemplo, utilizando norma 1 e tolerância 10^{-12}, o método Gauss-Seidel necessita 41829 iterações para fornecer a resposta. Sob as mesmas condições, o método de Jacobi necessita 110819 iterações. Se realizarmos um gráfico com o comportamento das normas, poderemos perceber que a mesma possui um caráter oscilante via método de Jacobi, enquanto que via método Gauss-Seidel o comportamento é monotônico.

6) $\alpha \in (-3, 3)$ e $\beta \in (-4, -3) \cup (3, 4)$.

7)

1. Se $|r| < 1$ e n ímpar, então $|\alpha| > 2\frac{r}{1-r}\left(1 - r^{\frac{n-1}{2}}\right)$.
2. Se $|r| < 1$ e n par, então $|\alpha| > 2\frac{r}{1-r}\left(1 - r^{\frac{n}{2}}\right) + r^{\frac{n}{2}-1}$.
3. Se $|r| > 1$, então $|\alpha| > \frac{r}{r-1}\left(r^{n-1} - 1\right)$.
4. Se $|r| = 1$, então $|\alpha| > n - 1$.

10) Aproximadamente $5{,}189 \times 10^{-12}$.

11) Aproximadamente $2{,}218 \times 10^{-12}$.

12) Aproximadamente 0.142105A entre os nós 1 e 4.

13) A reação com a menor taxa é a segunda, 0,391mol/(l s).

10.3 Capítulo 3

2) $x^* = -2{,}94753\ldots$ e $x^* = 1{,}50524\ldots$

3) A primeira equação possui raízes $x^* = -1$ (simples) e $x^* = 1{,}5$ (multiplicidade 3). A segunda equação possui raízes $x^* = -1$ (multiplicidade 3) e $x^* = 1{,}5$ (simples).

4) $x^* = 0{,}739085\ldots$

5) $x^* = 0{,}0946497\ldots$ e $x^* = 0{,}739533\ldots$

6) $x^* = 1{,}52430\ldots$; $x^* = 5{,}31204\ldots$ e $x^* = 6{,}72423\ldots$

8) No ponto x^*, a reta e a função assumem o mesmo valor, ou seja,

$$\alpha x^* = \text{sen} x^*.$$

E nesse mesmo ponto, a reta é tangente à função, ou seja,

$$\begin{aligned} \frac{d}{dx}(\alpha x)\Big|_{x=x^*} &= \frac{d}{dx}(\text{sen}x)\Big|_{x^*}, \\ &\Downarrow \\ \alpha &= \cos x^*. \end{aligned}$$

Combinando o valor de α em termos de x^* com a primeira equação, temos que x^* deve ser tal que

$$x^* \cos x^* = \text{sen} x^*.$$

A partir dessa última expressão, montamos a equação não linear $f(x^*) = 0$ e utilizamos o método. A solução é dada por

$$x^* = 7,7252\ldots$$

10) As raízes são -0.5 (simples), 0.5 (multiplicidade 3), $\sqrt{-1}$ e $-\sqrt{-1}$ (ambas simples).

11)

O gráfico dessa função permite visualizar que a localização do primeiro mínimo positivo esta na vizinhança próxima de $1,5$. Nesse mínimo, a derivada da função se anula, portanto estamos interessados no zero da função

$$f(x) := 3\,\frac{\cos 3x}{x} - \frac{\text{sen}3x}{x^2}$$

que está próximo de $1,5$. A melhor aproximação com seis dígitos é $1,49780$.

12)

O número que buscamos, $\sqrt[k]{y}$, é solução da equação $f(x) = 0$, onde

$$f(x) := x^k - y.$$

Assim, de acordo com o método Newton-Raphson, a relação de recorrência dada por

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

para $n = 0, 1, \ldots$ se aproxima do número $\sqrt[k]{y}$ para uma escolha de aproximação inicial $x^{(0)}$. No presente caso, a relação de recorrência para aproximar $\sqrt[8]{10}$ é dada por

$$x^{(n+1)} = \frac{7}{8}x^{(n)} + \frac{10}{8}\,\frac{1}{\left(x^{(n)}\right)^7}.$$

A convergência é quadrática pois a derivada de f em seu zero é $k\left(x^*\right)^{k-1}$, como $x^* \neq 0$, então $f'\left(x^*\right) \neq 0$.

16)

Podemos notar que o máximo global da função f está próxima do ponto $(-1.2, -1.3)$. No máximo (x^*, y^*), as derivadas parciais com relação às variáveis x e y são nulas, ou seja, o ponto de máximo é uma solução (mas não a única) do sistema

$$\left\{\begin{array}{rcl} \dfrac{\partial f}{\partial x}(x,y) & = & 0 \\ \\ \dfrac{\partial f}{\partial y}(x,y) & = & 0 \end{array}\right. \Longrightarrow \left\{\begin{array}{rcl} -4x^3 + 3y^3 - 1 & = & 0 \\ \\ -6y^5 + 9xy^2 & = & 0 \end{array}\right. . \tag{10.3.1}$$

No Scilab, podemos analisar o comportamento da função f através do gráfico para a superfície formada pelo conjunto de pontos $\left\{(x, y, f(x, y)) \in \mathbb{R}^3 | x, y \in \mathbb{R}\right\}$ e do gráfico para as curvas de nível dessa superfície. Começamos com a definição de uma função que dependa de dois argumentos escalares e retorne um escalar:

```
function z=fl(x,y)
  z=- x^4 - y^6 + 3*x*y^3 - x;
endfunction
```

Utilizaremos a função `fl` como argumento das instruções que montam os gráficos.

```
x=linspace(-1.5,1.5,50); // Definição do intervalo de valores nas
variáveis x e y.
y=linspace(-1.5,1.5,50);

fplot3d(x,y,fl); // Gráfico para a superfície.
contour(x,y,fl,-5:0.2:2.4,flag=[1 0 4],zlev=-35); // Gráfico para
as curvas de nível. Serão desenhadas as curvas para os níveis -5, -
4.8, -4.6, ..., 2.2 e 2.4 no plano z=-35. O argumento flag=[1 0 4]
corresponde a uma formatação particular do gráfico -- Consulte a ajuda
on-line do Scilab. Pode ser necessário alterar o ponto de vista para
melhor visualizar o gráfico.
```

A sequência de comandos acima desenha a superfície e as curvas de nível na mesma janela gráfica. É possível desenhá-las em janelas separadas. Nesse caso utilizamos o comando `scf()` e uma forma mais simples para as curvas de nível `contour(x,y,f2,-5:0.2:2.4)`.

A partir da informação sobre a aproximação inicial podemos criar uma função $F : \mathbb{R}^2 \to \mathbb{R}^2$ tal que a equação

$$F(t) = 0 \in \mathbb{R}^2$$

represente o sistema (10.3.1). Ou seja t será representado por um vetor coluna (uma matriz 2×1) cuja primeira componente, $t(1)$, corresponde à variável x e a segunda, $t(2)$, corresponde à y. No Scilab, definiremos F como a função `f2`:

```
function z=f2(t)
```

```
  z(1)=-4*t(1)^3 + 3*t(2)^3 -1;
  z(2)=-6*t(2)^5 + 9*t(1)*t(2)^2;
endfunction
```

Agora basta utilizá-la como argumento da função zero_newraph:

```
zero_newraph(f2,[-1.2;-1.3],100,1);
```

Observação 10.3.1. *1. O código zero_newraph deve ser carregado previamente.*

2. Note que a aproximação inicial é repassada como um vetor coluna.

3. A norma 1 *foi escolhida arbitrariamente.*

O resultado aproximado com seis dígitos é $(x^*; y^*) \approx (-1{,}15797; -1{,}20207)$.

17) Aproximadamente $5{,}49372 \times 10^{-3}$, ou seja, 0,549372% ao mês.

18) $[-4{,}21102 \times 10^{-1};\ -3{,}45229 \times 10^{-1}]$; $[0{,}45;\ -0{,}2]$; $[4{,}15590 \times 10^{-1};\ 3{,}773523 \times 10^{-1}]$ e $[2{,}53936 \times 10^{-1};\ 8{,}48516 \times 10^{-1}]$.

19) O máximo vale aproximadamente 0,698256.

20) Custo aproximado de 310,74 R$/MWh.

21) O semieixo maior da órbita vale aproximadamente 6831Km.

10.4 Capítulo 4

1) Se f admite uma expansão em série de Taylor em torno do ponto x, podemos estimar ao erro na operação de ponto flutuante:

$$\begin{aligned}
\left(D_{+,h}f\right)(x) &= \frac{f(x+h)-f(x)}{h} \\
&= \frac{f(x)+hf'(x)+\frac{h^2}{2}f''(x)+O(h^3)-f(x)}{h} \\
&= f'(x)+\frac{h}{2}f''(x)+O(h^2) = f'(x)+c_0h+O(h^2).
\end{aligned}$$

Assim, ao utilizarmos espaçamento $2h$, a operação será tal que

$$\left(D_{+,2h}f\right)(x) = f'(x)+c_0(2h)+O(h^2).$$

A seguinte combinação linear anula o termo com dependência linear em h:

$$\left(D_{+,2h}f\right)(x) - 2\left(D_{+,h}f\right)(x) = -f'(x)+O(h^2)$$

e a partir dela definimos a operação de diferença finita $\left(D_{+1,h}f\right)(x) := 2\left(D_{+,h}f\right)(x) - \left(D_{+,2h}f\right)(x) = f'(x)+O(h^2)$.

A nova operação $(D_{+1,h}f)(x)$ possui a seguinte dependência em h:

$$(D_{+1,h}f)(x) = f'(x) + c_1h^2 + O(h^3)$$

de modo que

$$(D_{+1,2h}f)(x) = f'(x) + c_1(2h)^2 + O(h^3).$$

E a combinação linear

$$4(D_{+1,h}f)(x) - (D_{+1,2h}f)(x) = 3f'(x) + O(h^3)$$

remove a dependência quadrática em h. Definimos $(D_{+2,h}f)(x)$ como

$$(D_{+2,h}f)(x) := \frac{4(D_{+1,h}f)(x) - (D_{+1,2h}f)(x)}{3},$$

ou seja,

$$\begin{aligned}(D_{+2,h}f)(x) &= \frac{4\left(2\frac{f(x+h)-f(x)}{h} - \frac{f(x+2h)-f(x)}{2h}\right) - \left(2\frac{f(x+2h)-f(x)}{2h} - \frac{f(x+4h)-f(x)}{4h}\right)}{3} \\ &= \frac{f(x+4h) - 12f(x+2h) - 32f(x+h) - 21f(x)}{12h}\end{aligned}$$

2) $f'(0{,}3) \approx -1{,}350$ e $f'(0{,}1) \approx 0{,}8833$

3) $O(h)$

4) $O(h)$

5) $\frac{-f(x-2h) + 16f(x-h) - 30f(x) + 16f(x+h) - f(x+2h)}{12h^2} = f''(x) + O\left(h^4\right)$

6) $\frac{f(x) - 2f(x+h) + f(x+2h)}{h} = f''(x) + O(h)$

7) $\frac{-3f(x) + 4f(x+h) - f(x+2h)}{2h} = f'(x) + O\left(h^2\right)$ e $\frac{f(x) - 2f(x+h) + f(x+2h)}{h} = f''(x) + O(h)$

8) O fluxo térmico vale aproximadamente $1{,}875 \times 10^4 \text{W/m}^2$

9) $\frac{f(x-4h) - 34f(x-2h) + 64f(x-h) - 64f(x+h) + 34f(x+2h) - f(x+4h)}{48h^3} = f'''(x) + O\left(h^4\right)$

10) $\frac{-4f(x-h) + 3f(x) + f(x+2h)}{6h} = f'(x) + O\left(h^2\right)$.

11) Uma possibilidade é (combinando Δ_h e Δ_{2h})
$\frac{8f(x-4h) - 64f(x-2h) + 40f(x+h) - 5f(x+2h) + 24f(x+3h) - 3f(x+6h)}{180h} = f'''(x) + O\left(h^3\right)$

12) $\frac{4y(x-3h) + (-16+h)y(x-2h) + 4(5-h)y(x-h) + (8-3h)y(x)}{2h^2} = y'(x) - 2y''(x) + O\left(h^2\right)$

13) $T'(R) \approx \dfrac{9T(R) - 12T\left(\frac{5}{6}R\right) + 3T\left(\frac{4}{6}R\right)}{R}$

10.5 Capítulo 5

2) A partir dos dados da tabela encontramos as seguintes interpolações. Observação: os resultados foram obtidos a partir de operações em ponto flutuante e os primeiros sete dígitos estão representados. Se admitirmos os valores da tabela como racionais exatos, a interpolação envolverá apenas coeficientes racionais. Por exemplo, a interpolação da função seno será simplesmente $P(x) = x$; nesse caso, a diferença deve-se exclusivamente aos erros de arredondamento cometidos nas operações aritméticas.

para a função cotangente $P_{cot}(x) = 2283{,}332\ldots - 1{,}874997\ldots \cdot 10^6 x + 7{,}083307\ldots \cdot 10^8 x^2 - 1{,}249991\ldots \cdot 10^{11} x^3 + 8{,}33325 \cdot 10^{12} x^4$.

para a função seno $P_{sen}(x) = -3{,}469446\ldots \cdot 10^{-18} + x + 7{,}275957\ldots \cdot 10^{-12} x^2 - 1{,}862451\ldots \cdot 10^{-9} x^3$.

para a função cosseno $P_{cos}(x) = 1{,}000008\ldots - 0{,}01399999\, x + 7{,}833333\, x^2 - 2000\, x^3 + 166666{,}6\, x^4$.

O erro cometido na aproximação do valor de $\cot(0{,}0015)$ por $\dfrac{P_{cos}(0{,}0015)}{P_{sen}(0{,}0015)}$ pode ser avaliada através da propagação de erros. De acordo com ela, o erro – vamos denominá-lo $\delta\dfrac{P_{cos}}{P_{sen}}(x)$ – está relacionado aos erros cometidos na aproximação do seno e do cosseno pelas respectivas interpolações[2]:

$$\delta\frac{P_{cos}}{P_{sen}}(x) \approx \left|\frac{1}{P_{sen}(x)}\right| \delta P_{cos}(x) + \left|\frac{P_{cos}(x)}{(P_{sen}(x))^2}\right| \delta P_{sen}(x),$$

onde os erros são dados por $\delta P_{cos}(x) = |\cos(x) - P_{cos}(x)|$ e $\delta P_{sen}(x) = |\text{sen}(x) - P_{sen}(x)|$.

Dessa forma, $\delta P_{cos}(0{,}0015) = 1{,}5625\ldots \cdot 10^{-7}$ e $\delta P_{sen}(0{,}0015) = 5{,}625\ldots \cdot 10^{-10}$ que implicam a estimativa $\delta\dfrac{P_{cos}}{P_{sen}}(0{,}0015) \approx 0{,}0694449\ldots$. Já o erro obtido a partir da interpolação direta da cotangente é $\delta P_{cot}(0{,}0015) = |\cot(0{,}0015) - P_{cot}(0{,}0015)| = 18{,}2292\ldots$.

A diferença entre as duas estimativas se deve ao fato de que a função cotangente varia muito rapidamente no intervalo de pontos escolhido, o que potencialmente aumenta os erros de truncamento. Por outro lado, as funções seno e cosseno variam pouco, o que permite uma interpolação com menos erro de truncamento.

3) Devemos construir duas interpolações polinomiais, uma com 4 pontos igualmente espaçados, $x_j = \dfrac{1}{3}(j-1)$, para $j = 1, 2, 3, 4$ e a outra com pontos espaçados segundo a fórmula de Chebishev, $x_j = \dfrac{a+b}{2} + \dfrac{(a-b)}{2}\cos\left(\dfrac{2j-1}{2n}\pi\right)$, com $n = 4$ e $j = 1, 2, 3, 4$.

- A função $\text{sen}(2x)$ possui expansão em série de Taylor em torno de $x = 0.5$ dada por $\text{sen}(2x) = 0{,}8414709\ldots + 1{,}080604\ldots(x - 0{,}5 - 1{,}682941\ldots(x - 0{,}5)^2 - 0{,}7204037\ldots(x -$

[2] Calculamos a cotangente através de uma expressão do tipo $f(z_1, z_2) = \dfrac{z_1}{z_2}$, portanto, de acordo com a expressão para propagação de erros, $\delta f(z_1, z_2) \approx \left|\dfrac{\partial f}{\partial z_1}(z_1, z_2)\right| \delta z_1 + \left|\dfrac{\partial f}{\partial z_2}(z_1, z_2)\right| \delta z_2$.

$0{,}5)^3+O((x-0{,}5)^4)$. A interpolação com pontos igualmente espaçados é $P(x) = 2{,}10091\,x - 0{,}510277\,x^2 - 0{,}681331\,x^3$. A interpolação com os pontos de Chebyshev é $P(x) = -0{,}00356134 + 2{,}11239\,x - 0{,}519039\,x^2 - 0{,}685124\,x^3$.

- A função e^x possui expansão em série de Taylor em torno de $x = 0{,}5$ dada por $e^x = 1{,}64872 + 1{,}64872(x-0{,}5) + 0{,}824361(x-0{,}5)^2 + 0{,}274787(x-0{,}5)^3 + O((x-0{,}5)^4)$. A interpolação com pontos igualmente espaçados é $P(x) = 1 + 1{,}01399\,x + 0{,}425665\,x^2 + 0{,}278626\,x^3$. A interpolação com os pontos de Chebyshev é $P(x) = 0{,}999509 + 1{,}01563\,x + 0{,}424301\,x^2 - 0{,}27824\,x^3$.

- A função $\sqrt{x}$ possui expansão em série de Taylor em torno de $x = 0{,}5$ dada por $\sqrt{x} = 0{,}707107 + 0{,}707107(x-0{,}5) + 0{,}353553(x-0{,}5)^2 + 0{,}353553(x-0{,}5)^3 + O((x-0{,}5)^4)$. A interpolação com pontos igualmente espaçados é $P(x) = 2{,}52192\,x - 2{,}79344\,x^2 + 1{,}27152\,x^3$. A interpolação com os pontos de Chebyshev é $P(x) = 0{,}127449 + 1{,}8407\,x - 1{,}69585\,x^2 + 0{,}732394\,x^3$.

- A função $\dfrac{1}{1+25x^2}$ possui expansão em série de Taylor em torno de $x = 0.5$ dada por $\dfrac{1}{1+25x^2} = 0{,}137931 - 0{,}475624(x-0{,}5) + 1{,}16446(x-0{,}5)^2 - 2{,}37529(x-0{,}5)^3 + O((x-0{,}5)^4)$. A interpolação com pontos igualmente espaçados é $P(x) = 1 - 3{,}45075\,x + 4{,}35728\,x^2 - 1{,}86807\,x^3$. A interpolação com os pontos de Chebyshev é $P(x) = 1{,}11311 - 4{,}09332\,x + 5{,}43048\,x^2 - 2{,}42567\,x^3$.

- A função x^4 possui expansão em série de Taylor em torno de $x = 0{,}5$ dada por $x^4 = -0{,}0625 + 0{,}5(x-0{,}5) + 1{,}5(x-0{,}5)^2 + 2(x-0{,}5)^3 + O((x-0{,}5)^4)$. A interpolação com pontos igualmente espaçados é $P(x) = 0{,}222222\ldots x - 1{,}222222\ldots\, x^2 + 2\,x^3$. A interpolação com os pontos de Chebyshev é $P(x) = -0{,}0078125 + 0{,}25\,x - 1{,}25\,x^2 + 2\,x^3$.

5) O spline cúbico natural para o conjunto de dados formado por seis pontos é construído a partir de cinco polinômios de grau 3: $s_i(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3$, onde $i = 1, 2, \ldots, 5$ mais o coeficiente acessório c_6. Como trata-se de um spline natural $c_1 = c_6 = 0$, os demais coeficientes c_i são solução solução do sistema

$$\begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 \\ -6 \\ 0 \\ 6 \end{pmatrix}.$$

A solução é o vetor $(0{,}401914, -1{,}60766, 0{,}0287081, 1{,}49282)^T$. Os coeficientes b_i e d_i são calculados a partir de $c_1, c_2 \ldots c_5$, os coeficientes a_i são calculados a partir da relação $a_i = f_i$.

6) Para que uma função $g(x)$ seja um spline de ordem n é necessário que as derivadas de ordem $0 \leq k \leq n-1$ dos polinômios que a constitui, $s_i(x)$, se igualem nos pontos de interpolação, ou seja $s_i^{(k)}(x_{i+1}) = s_{i+1}^{(k)}(x_{i+1})$:

1. $f(x) = \begin{cases} x & , \quad -1 \leq x < 0 \\ 2x & , \quad 0 \leq x < 1 \\ x+1 & , \quad 1 \leq x \leq 2 \end{cases}$. Os pontos de interpolação internos são $x = 0$ e $x = 1$, de acordo com a expressão para f, os limites nesses pontos existem: $\lim_{x\to 0} f(x) = 0$ e $\lim_{x\to 1} f(x) = 2$. Quanto à derivada, $f'(x) = \begin{cases} 1 & , \quad -1 < x < 0 \\ 2 & , \quad 0 < x < 1 \\ 1 & , \quad 1 < x < 2 \end{cases}$, ou seja, a derivada não é contínua. Portanto , como apenas f é contínua, essa função é um spline linear.

2. $f(x) = \begin{cases} x & , \quad -1 \leq x < 0 \\ 2x-1 & , \quad 0 \leq x < 1 \\ x+1 & , \quad 1 \leq x \leq 2 \end{cases}$.Os pontos de interpolação internos são $x = 0$ e $x = 1$, de acordo com a expressão para f, o limite no ponto $x = 0$ não existe, pois $\lim_{x\to 0^-} f(x) = 0$ e $\lim_{x\to 0^+} f(x) = -1$. Isto é o suficiente para garantir que f não é um *spline*.

3. $f(x) = \begin{cases} 0 & , \quad -1 \leq x < 0 \\ x^2 & , \quad 0 \leq x < 1 \\ 2x-1 & , \quad 1 \leq x \leq 2 \end{cases}$. Os pontos de interpolação internos são $x = 0$ e $x = 1$, de acordo com a expressão para f, os limites nesses pontos existem: $\lim_{x\to 0} f(x) = 0$ e $\lim_{x\to 1} f(x) = 1$. Quanto à derivada, $f'(x) = \begin{cases} 0 & , \quad -1 < x < 0 \\ 2x & , \quad 0 < x < 1 \\ 2 & , \quad 1 < x < 2 \end{cases}$, podemos verificar que os limites em $x = 0$ e $x = 1$ existem: $\lim_{x\to 0} f'(x) = 0$ e $\lim_{x\to 1} f(x) = 2$. Então, como f envolve apenas polinômios de grau menor ou igual a 2, segue que f é um *spline* quadrático.

7) Dada a função

$$f(x) = \begin{cases} x^3 + x & , \quad -1 \leq x \leq 0 \\ & \\ ax^2 + bx & , \quad 0 \leq x \leq 1 \end{cases}, \tag{10.5.1}$$

temos que

$$f'(x) = \begin{cases} 3x^2 + 1 & , \quad -1 < x < 0 \\ & \\ 2ax + b & , \quad 0 < x < 1 \end{cases} \tag{10.5.2}$$

e

$$f''(x) = \begin{cases} 6x & , \quad -1 < x < 0 \\ & \\ 2a & , \quad 0 < x < 1 \end{cases}. \tag{10.5.3}$$

Para que f seja um *spline* cúbico, ela deve ser tal que os limites para f, f' e f'' devem estar bem definidos em $x = 0$. Ou seja, a partir de (10.5.3) , devemos ter que $a = 0$; a partir de (10.5.2), devemos ter que $b = 1$. Quaisquer que sejam os valores de a e b, $\lim_{x\to 0} f(x) = 0$. Portanto, f será um *spline* cúbico se $a = 0$ e $b = 1$.

8) Será um spline cúbico se $c = -9$ e $b = 1$. Nunca será um spline cúbico natural, pois a

condição necessária $c = -9$ implica $s''(1) \neq 0$.

9) $x \approx 2{,}35028$

10) Aproximadamente $-7{,}47$°C

11) Aproximadamente $-38{,}6$°C

12) Aproximadamente 0,99953L

13) O valor máximo para l/l_0 é aproximadamente 1,00051 (a 200K) e o mínimo 0,999819 (a 500K).

14) Aproximadamente 0,176m/s (ocorre em $t \approx 0{,}8315323$)

15) Aproximadamente 0,844 unidades de tempo.

16) Aproximadamente 0,1245 unidades de comprimento (ocorre em $t \approx 0{,}411392$ unidades de tempo).

10.6 Capítulo 6

1) $a \approx 6{,}37$ e $b \approx 0{,}950$

2) $t_{25} \approx 0{,}99061$h.

3) O valor máximo que a função f ajustada assume é aproximadamente 0,5043

4) O erro relativo é de aproximadamente $0{,}2140 \times 10^{-2}$ para o coeficiente a_1 e $0{,}3925 \times 10^{-2}$ para o coeficiente a_2.

5) A soma do quadrado dos resíduos para a função modelo ajustada φ vale aproximadamente 0.491×10^{-2}, enquanto que no caso da função modelo ajustada ψ a soma vale aproximadamente $0{,}.353 \times 10^{-1}$. Portanto, a função modelo que mais se ajusta aos dados é a função φ.

6) $A \approx 0{,}07728$, $k \approx 3{,}577$, $b \approx 1{,}030$ e $c \approx -0{,}2146$

7) $A \approx 8{,}213$, $b \approx 6{,}664$ e $c^2 \approx 0{,}1716$

8) O semieixo maior da órbita é igual a $2A$. De acordo com o ajuste, esse valor é de aproximadamente $1{,}251 \times 10^4$km.

10.7 Capítulo 7

1)

1.1) Basta checar os monômios, x^n, $n = 0, 1, \ldots$ A integração é exata apenas para $n = 0$ e 1.

1.2) Não é uma regra obtida a partir de uma interpolação polinomial. Se o fosse, seria exata para polinômios de grau 2 e de acordo com o item anterior, ela não é exata nesse caso. $C_1 f(0{,}2) + C_2 f(0{,}5) + C_3 f(0{,}8)$, onde $C_1 = \dfrac{25}{54}$, $C_2 = \dfrac{2}{27}$ e $C_3 = \dfrac{25}{54}$.

1.3) Devemos realizar uma mudança de variável, na forma de uma transformação afim, para passar do intervalo de integração $(1,3)$ para o intervalo $(0,1)$ onde a regra está definida. A transformação deve ser da forma $y = \frac{1}{2}x - \frac{1}{2}$, ou ainda, $x = 2y+1$. O Jacobiano dessa transformação vale 2: $dx = 2dy$. Assim, os pesos da regra original serão multiplicados por 2.

2) De acordo com a regra composta de Simpson,

$$\int_a^b f(x)\,dx = \frac{h}{3}\sum_{k=1}^{n} C_k f(x_k) - \frac{h^4}{180}(b-a)f^{(4)}(\xi).$$

Dessa forma, o erro de truncamento está relacionado aos valores que assume a derivada de quarta ordem do integrando.

$$\frac{d^4}{dx^4}\left(e^{x^2}\right) = 4e^{x^2}\left(3 + 12x^2 + 4x^4\right).$$

Neste caso, o valores mínimo e o máximo da derivada ocorrem respectivamente nos extremos inferior e superior do intervalo $(0,1)$, portanto

$$12 \leq \left.\frac{d^4}{dx^4}\left(e^{x^2}\right)\right|_{x=\xi} \leq 76e \approx 206{,}6$$

Dessa forma, o erro de truncamento está contido no intervalo

$$-\frac{207h^4}{180} \leq \int_0^1 e^{x^2}\,dx - \frac{h}{3}\sum_{k=1}^{n} C_k e^{x_k^2} \leq -\frac{12h^4}{180}.$$

Se desejarmos que o erro de truncamento em valor absoluto seja menor ou igual a 10^{-6}, então h deve estar no intervalo

$$0{,}0305369 \approx \left(\frac{180}{207}10^{-6}\right)^{\frac{1}{4}} \leq h \leq \left(\frac{180}{12}10^{-6}\right)^{\frac{1}{4}} \approx 0{,}0622333.$$

3) O valor exato da integral é $\dfrac{tg^{-1}5}{5} \approx 0{,}27468$.
Newton-Cotes.

- 3 pontos (regra de Simpson): $\approx 0{,}2650309$
- 4 pontos (regra 3/8): $\approx 0{,}2600357$
- 5 pontos (regra de Boole): $\approx 0{,}2615188$

Quadratura por interpolação em n pontos de Chebishev, $x_i = \frac{a+b}{2} + \frac{a-b}{2}\cos\left(\frac{2i-1}{2n}\pi\right)$, $i = 1, 2, \ldots, n$.

- 3 pontos: $\approx 0{,}2861982$
- 4pontos: $\approx 0{,}2701897$
- 5 pontos: $\approx 0{,}2714549$

Quadratura gaussiana.

- 3 pontos: $\approx 0{,}2855636$
- 4 pontos: $\approx 0{,}2771697$
- 5pontos: $\approx 0{,}2743215$

Romberg.

Nesse caso, 5 pontos correspondem à utilização de duas quadraturas compostas, $5 = 2^2 + 1$. Assim,

$$R(2,2) \approx 0{,}2615188$$

4) A distância total percorrida é dada pela integral

$$\int_0^{50} \sqrt{v_x(t)^2 + v_y(t)^2}dt$$

e a posição final é dada pelo ponto (x, y), onde

$$x = \int_0^{50} v_x(t)dt \qquad \text{e} \qquad y = \int_0^{50} v_y(t)dt.$$

As integrais devem ser calculadas de acordo com a regra composta de Simpson:

$$\int_0^{50} \sqrt{v_x(t)^2 + v_y(t)^2}dt \approx \sum_{i=1}^{11} C_i \sqrt{vx_i^2 + vy_i^2},$$

$$\int_0^{50} v_x(t)dt \approx \sum_{i=1}^{11} C_i\, vx_i$$

e

$$\int_0^{50} v_y(t)dt \approx \sum_{i=1}^{11} C_i\, vy_i.$$

A partir das aproximações temos que a posição final é dada por $\approx (35{,}4; 17{,}717)$ e o deslocamento total é de $\approx 56{,}75$.

5) A quadratura será exata para qualquer polinômio de grau menor ou igual a quatro.

$$\begin{cases} C_1 + C_2 + C_3 &= 2 \\ \alpha C_1 + \frac{1}{3}C_2 + \omega C_3 &= 0 \\ \alpha^2 C_1 + \frac{1}{3^2}C_2 + \omega^2 C_3 &= 2/3 \\ \alpha^3 C_1 + \frac{1}{3^3}C_2 + \omega^3 C_3 &= 0 \\ \alpha^4 C_1 + \frac{1}{3^4}C_2 + \omega^4 C_3 &= 2/5 \end{cases}$$

6)

$$1{,}7709999999969849 \approx 1.771 - \frac{60{,}3}{2\times 10^{13}} \leq \int_1^{10} f(x)dx \leq 1{,}771 + \frac{22{,}1}{2\times 10^{13}} \approx 1{,}7710000000011049$$

7) $I \approx 6{,}18274 \times 10^{-3}$.

8) $I \approx -0{,}0497840$.

9) Aproximadamente 401m³/s.

10) Aproximadamente 8839m³

10.8 Capítulo 8

1) Velocidade terminal, v_T, é aquela que implica $\frac{dv}{dt}(v_T) = 10 - 0{,}00343v_T^2 = 0$, ou seja, $v_T = 53{,}9949\ldots$.De acordo com o método R-K de 4 estágios, aproximamos a solução v nos pontos t_i por $v_i \approx v(t_i)$, para $i = 0, 1, 2, \ldots$

Inicialmente realizamos a escolha, $t_i = ih$ com $h = 1$. De acordo com o método

$$v_{i+1} = v_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

onde

$$\begin{aligned} k_1 &= 10 - 0{,}00343v_i^2, \\ k_2 &= 10 - 0{,}00343\left(v_i + \frac{h}{2}k_1\right)^2, \\ k_3 &= 10 - 0{,}00343\left(v_i + \frac{h}{2}k_2\right)^2, \\ k_4 &= 10 - 0{,}00343\left(v_i + hk_3\right)^2 \end{aligned}$$

e $v_0 = v(0) = 0$. A escolha $h = 1$ implica

$$\begin{aligned} v_1 &= 9{,}88711\ldots, \\ v_2 &= 19{,}1326\ldots, \\ v_3 &= 27{,}2514\ldots, \\ v_4 &= 33{,}9963\ldots \end{aligned}$$

Como $\frac{1}{2}v_T = 26{,}9975\ldots$ a aproximação $v_3 \approx v(t_3)$ está próxima do valor exato. De acordo com a escolha $h = 1$, $t_3 = 3$. Se as aproximações forem refeitas com $h = 0{,}5$ notaremos que as aproximações obtidas com espaçamento $h = 1$. são exatas para os primeiros dígitos.

2) Por meio das novas variáveis $z_1(t) := \theta(t)$ e $z_2(t) = \frac{d\theta}{dt}(t)$ reescrevemos a E.D.O. de segunda

ordem como o seguinte sistema de E.D.O de primeira ordem:

$$\begin{cases} \dfrac{dz_1}{dt} &= z_2 \\ \dfrac{dz_2}{dt} &= -\dfrac{g}{l}\text{sen}(z_1) \end{cases}$$

com condição inicial $z_1(0) = \dfrac{\pi}{4}$ e $z_2(0) = 0$. De acordo com o método Runge-Kutta de 4 estágios, a aproximação $z_{1,i} \approx \theta(t_i)$, $z_{2,i} \approx \dfrac{d\theta}{dt}(t_i)$ é determinada pelo conjunto de equações

$$\begin{cases} z_{1,i+1} &= z_{1,i} + \dfrac{h}{6}\left(k_{1,z_1} + 2k_{2,z_1} + 2k_{3,z_1} + k_{4,z_1}\right) \\ z_{2,i+1} &= z_{2,i} + \dfrac{h}{6}\left(k_{1,z_2} + 2k_{2,z_2} + 2k_{3,z_2} + k_{4,z_2}\right) \end{cases},$$

onde

$$\begin{aligned} k_{1,z_1} &= z_{2,i}, & k_{1,z_2} &= -\frac{g}{l}\text{sen}\left(z_{1,i}\right), \\ k_{2,z_1} &= z_{2,i} + \frac{h}{2}k_{1,z_2}, & k_{2,z_2} &= -\frac{g}{l}\text{sen}\left(z_{1,i} + \frac{h}{2}k_{1,z_1}\right), \\ k_{3,z_1} &= z_{2,i} + \frac{h}{2}k_{2,z_2}, & k_{3,z_2} &= -\frac{g}{l}\text{sen}\left(z_{1,i} + \frac{h}{2}k_{2,z_1}\right), \\ k_{4,z_1} &= z_{2,i} + hk_{3,z_2}, & k_{4,z_2} &= -\frac{g}{l}\text{sen}\left(z_{1,i} + hk_{3,z_1}\right), \end{aligned}$$

e $z_{1,0} = \dfrac{\pi}{4}$, $z_{2,0} = 0$.

No caso em que $g = 10$ e $l = 1$, a escolha $t_i = ih$ com $h = 0{,}1$ determina as aproximações

t_i	θ_i	θ'_i
0,1	7,502529881968854619D-01	-6,986640560618003759D-01
0,2	6,473769250333341052D-01	-1,345764731270578274D+00
0,3	4,846901251555982282D-01	-1,885214951935509031D+00
0,4	2,758968318499587791D-01	-2,258596094811123667D+00
0,5	4,020089739876930857D-02	-2,416906369014381273D+00
0,6	-1,994576486681848282D-01	-2,336869344902609225D+00
0,7	-4,195523688530167772D-01	-2,030454605693730485D+00
0,8	-5,993625050307915814D-01	-1,539848623569317176D+00
0,9	-7,232646149718042761D-01	-9,220589076791289029D-01
1,0	-7,814502189975390811D-01	-2,346645475407357351D-01

t_i	$\frac{E_i}{m}$
0,1	-7,07109825738047170 8D+00
0,2	-7,07114371708521005 6D+00
0,3	-7,07117602124684996 3D+00
0,4	-7,07118462701447292 3D+00
0,5	-7,07120232914960666 9D+00
0,6	-7,07126265514007190 7D+00
0,7	-7,07134079816888494 5D+00
0,8	-7,07138714737948426 1D+00
0,9	-7,07139458170268309 1D+00
1,0	-7,07139509128359922 1D+00

3) Soluções aproximadas:

t_i	θ_i	θ'_i
0,1	7,51272653925855626 9D-01	-6,59355490485860507 0D-01
0,2	6,61887615712736798 7D-01	-1,07726020799735300 3D+00
0,3	5,46278794120964361 6D-01	-1,19080166315519853 1D+00
0,4	4,30375248524299425 2D-01	-1,10639027518563581 9D+00
0,5	3,27155064967939834 8D-01	-9,54461209425327972 2D-01
0,6	2,39210752249669034 2D-01	-8,08163133677632927 7D-01
0,7	1,64482989459336870 2D-01	-6,91743559125479468 0D-01
0,8	9,97765509488913460 2D-02	-6,07320739502650308 6D-01
0,9	4,20883121757300096 6D-02	-5,50566774607573195 0D-01
1,0	-1,11066162317466701 8D-02	-5,16781875081264829 9D-01

t_i	l_i	l'_i
0,1	1,03567281506745567 7D+00	7,18892525814716854 0D-01
0,2	1,14542004997515789 5D+00	1,48166019449896602 1D+00
0,3	1,33188273286556357 9D+00	2,23658898604398892 5D+00
0,4	1,58832094923044531 1D+00	2,85970457718661563 7D+00
0,5	1,89544479116716813 0D+00	3,23485825409380733 9D+00
0,6	2,22464538033761316 5D+00	3,29443441539016435 0D+00
0,7	2,54311209823465889 1D+00	3,02099557102397131 6D+00
0,8	2,81845227509348061 9D+00	2,43894117155520762 1D+00
0,9	3,02246272297130902 4D+00	1,60635644277990330 4D+00
1,0	3,13409850737801409 8D+00	6,06909612533648501 5D-01

t_i	$\frac{E_i}{m}$
0,1	- 1,414197206858997635D+00
0,2	- 1,414157988548052414D+00
0,3	- 1,414169370386709534D+00
0,4	- 1,414237661239337696D+00
0,5	- 1,414275527535049815D+00
0,6	- 1,414276377918547878D+00
0,7	- 1,414267862850715041D+00
0,8	- 1,414263795974782134D+00
0,9	- 1,414266517191248518D+00
1,0	- 1,414274948146285737D+00

4)Soluções aproximadas:

t_i	$\theta 1_i$	$\theta 1'_i$
0,1	7,504685422015303642D-01	-6,897770467520577542D-01
0,2	6,510496003775438911D-01	-1,270930782712977436D+00
0,3	5,037684879639487967D-01	-1,631803386098856556D+00
0,4	3,341536192802013749D-01	-1,715011247000679750D+00
0,5	1,697120647693060036D-01	-1,535981627624850798D+00
0,6	3,311457155810154651D-02	-1,183559404914650948D+00
0,7	-6,917259335325923186D-02	-9,019594905497929638D-01
0,8	-1,592353515323187696D-01	-9,499724170982495330D-01
0,9	-2,652610976057471759D-01	-1,171112655424180415D+00
1,0	-3,898149186659949916D-01	-1,291058696342471723D+00

t_i	$\theta 2_i$	$\theta 2'_i$
0,1	7,849669880187599702D-01	-1,778069435585754102D-02
0,2	7,779841968635019533D-01	-1,522880644673337558D-01
0,3	7,461311401065024995D-01	-5,309885715954084651D-01
0,4	6,622435794163086253D-01	-1,184591763148334476D+00
0,5	5,035887915655447022D-01	-1,998928228485088887D+00
0,6	2,635726284306724176D-01	-2,770230361480222392D+00
0,7	-3,848503433780764427D-02	-3,178385013731055864D+00
0,8	-3,490616349961885856D-01	-2,930037024491089159D+00
0,9	-6,087310308453142138D-01	-2,221655003378214310D+00
1,0	-7,899621003083197035D-01	-1,404442674935349800D+00

t_i	$\frac{E_i}{m}$
0,1	-2,021329639777180631D+01
0,2	-2,02134094749300246OD+01
0,3	-2,021369606615465386D+01
0,4	-2,021379661120513660D+01
0,5	-2,021336466750859273D+01
0,6	-2,021373921783010630D+01
0,7	-2,021474467827373189D+01
0,8	-2,021545565083874152D+01
0,9	-2,021500305318189561D+01
1,0	-2,021526803314096554D+01

5) Aproximação com seis dígitos para o período: 1,89717.

6) 252,86.

7) Aproximação com seis dígitos para o ângulo: 0,0237355rad.

8) Amplitude pico a pico vale aproximadamente 4,457.

9)

$$x_{i+1} = x_i + \frac{h}{2}\left(K_{1,x} + K_{2,x}\right), \qquad y_{i+1} = y_i + \frac{h}{2}\left(K_{1,y} + K_{2,y}\right),$$

onde

$$K_{1,x} = t_i + x_i\, y_i, \qquad K_{1,y} = y_i + t_i\, x_i,$$

$$K_{2,x} = t_i + h + \left(x_i + h\, K_{1,x}\right)\left(y_i + h\, K_{1,y}\right), \quad K_{2,y} = y_i + h\, K_{1,y} + \left(t_i + h\right)\left(x_i + h\, K_{1,x}\right).$$

10) Potencial ação.

11) 152,0s

12) Valor mínimo para $x \approx 0{,}5965$. Valor mínimo para $y \approx 0{,}4688$.

13) Aproximação a partir de 101 pontos: 12,738772. Aproximação a partir de 201 pontos: 12,738746. Aproximação a partir de 401 pontos: 12,738744. O conjunto dessas aproximações permite concluir que a integral vale aproximadamente 12,73874.

14) O ângulo vale aproximadamente $-0{,}1057$ radianos, ou seja, $-6{,}055°$.

Apêndice

Teorema 10.8.1 (Valor Médio)
Seja $f : [a, b] \to \mathbb{R}$ *uma função contínua e diferenciável no intervalo aberto* (a, b)*, então existe um* $c \in (a, b)$ *tal que*

$$f(b) - f(a) = f'(c)(b - a).$$